Journal of Ubiquitous Systems & Pervasive Networks
Volume 1, No. 1 (2010) pp. 29-37

Adaptive Scheduling of Message Carrying in a Pigeon Network **

Jiazhen Zhou [a] *, Jiang Li [a] , Kenneth Mitchell [b]

[a] *Department of Systems and Computer Science Howard University, Washington DC, 20059*
[b] *Department of Computer Science Electrical Engineering University of Missouri -Kansas City 64110*

Abstract

In adverse environments where real-time communications are not always available, a delay tolerant network may be the only choice. A special type of delay tolerant network known as a ``pigeon network" utilizes controllable special purpose vehicles called ``pigeons" to convey messages among segregated areas. A challenging problem studied in this paper is how to schedule a ``pigeon" to enter/leave a particular segregated area in an optimal way so that the average delay of messages is minimized. This problem falls in the range of server vacation queueing model. Unfortunately, the amount of work in the open literature that addresses the optimal scheduling (service discipline) problem is very limited, especially when bursty traffic is considered. Based on an analysis of a scenario with deterministic and Poisson arrivals, we propose an adaptive scheduling scheme that alternates the service disciplines according to the trend of message accumulation on the home/foreign host. This adaptive scheme shows good performance and flexibility in adjusting to the dynamics of bursty and non-bursty traffic.

Keywords: *Disruption/Delay-tolerant networking, pigeon networks, scheduling, service discipline*

1. Introduction

Computer networks have found applications in more and more complex and diverse environments, including adverse ones that real-time communications are not always available. For example, when disasters such as earthquakes damage the communication infrastructure, a network becomes partitioned and it is hard to maintain real-time communications. A disruption/delay tolerant network (DTN) [1] is one of the effective technologies that support communications under these adverse environments.

Pigeon networks [10], which borrows the ancient idea of employing pigeons as the communication tool, can be viewed as a special type of DTN that use special-purpose message carriers. Of course, the “pigeons” used here are not the real pigeons. Instead, they are vehicles that are equipped with much better moving ability and partial instant wireless communication ability. For instance, it can be an unmanned aviation vehicle or a robotic insect. Similar to a real-life pigeon, the ``pigeon" in a pigeon network only serves its owner, which is denoted as its ``home host". Correspondingly, the other nodes that are not the owner are denoted as ``foreign hosts" with regard to this pigeon. In a pigeon network, the “pigeon” is used to convey messages between the home host and the foreign host.

The main goal of this paper is to explore strategies to schedule the pigeon so that the average delay of messages generated on the home host and foreign host is minimized. If we view the time that the message carrier is away from the foreign host (for delivery) as a vacation, the problem presented in this paper can be modeled by a queue with server vacations. However, a main difference from the regular vacation model is that the “service”

* Corresponding author. E-mail: zhouj@networks.howard.edu
** The work was funded in part by NSF grant CNS-0832000 and the Mordecai Wyatt Johnson Program of Howard University.

DOI: 10.5383/JUSPN.01.01.004

of each message includes two stages: pickup and delivery. This makes the definition of delay also different. Optimization for this special vacation model essentially turns into finding the optimal service discipline, which has not been fully addressed in the literature, especially when the bursty traffic is considered. The main contributions of this paper include: (1) The development of analytic models that include bursty traffic; (2) An algorithm for scheduling message carriers according to the dynamic change of arrivals. This algorithm adapts to the dynamics of the traffic. It is easy to implement and can guarantee close-to-optimal performance for any practical situation, as it does not require any a priori knowledge about the distribution or average rate of arrivals.

The main content of this paper begins with the introduction of related work in Section 2 and a description of the basic model in Section 3. To explore the optimal or suboptimal service discipline that can work well in an environment with dynamic traffic, analysis of basic arrival distributions such as the deterministic, Poisson, and periodic bursty distributions are presented in Section 4 and Section 5. The knowledge acquired about different service disciplines is then applied toward configuring an adaptive scheduling algorithm for general arrivals (Section 6). In this adaptive scheme, the instant that a trip ends (where service disciplines make a difference) is determined according to the trend of message accumulation on the home/foreign host. The numerical results show that it is superior to any monotonic scheme such as gated or exhaustive scheduling.

2. Related Work

2.1. Service disciplines in queueing model with server vacation

Queues with server vacations have been widely studied since the 1970's, e.g. work by Eisenberg [2], Levy and Yelichali [3], Fuhrmann and Cooper [4], and Doshi [5]. The main idea is that a server might need to be away from serving a queue for a certain time, either for taking a rest or serving other queues. The moment that a server begins and ends the service is determined by the service discipline employed. The main service disciplines studied include the *exhaustive service discipline*, the *gated service discipline*, and different kinds of threshold based, or *limited-K service disciplines*.

For an exhaustive service discipline, the server keeps serving the queue till it becomes empty. With a gated service discipline, the server only serves tasks seen upon its return from a vacation. The limited-K service discipline is a variation of the exhaustive or gated service discipline in which the server will not serve more than K tasks in a single cycle. Although different service disciplines have been studied with the average delay of each task being the main focus for the vacation model, there is no discussion about which service discipline is the best. This is because each task is viewed as being finished after it is served, which leads to the fact that the exhaustive service discipline is considered the best policy to achieve the minimum average delay. When optimal service disciplines need to be considered, the study of these systems usually involve the optimal control of queues.

2.2. Optimal control of queues

The main objective of the optimal control of queues (actually it is accomplished through control of the server and arrivals) is to minimize the running cost of a queue. Among the different factors that are considered are: the delay that each task suffers, the length of time that a server can rest or be away serving another queue, and the cost to keep a dedicated server [6].

The control of the server is usually accomplished in two ways: one is to control when the server starts and ends a service cycle, another is to control the service rate or number of servers that can be used for a single queue [7]. The rule of starting or ending service is basically the definition of a service discipline such as the vacation model. The threshold-based policy, where the server turns on when there are at least K tasks in the queue and then serves until the queue empties, has been proven to be the optimal control policy for a number of different applications [8]. As an alternative to determining the optimal moment of the start of a service or departure for a vacation, the number of servers can be increased or decreased according to the arrival rate and the length of the queue [9].

The main difference of our work is that the main objective of the optimization is the average delay that each message suffers. As we will demonstrate in this paper, the exhaustive service might not be the best for the pickup and delivery scenario (actually, for some cases, it might be the worst). The idea that a server will not start service until the queue reaches a certain length in order to improve the efficiency (see Zhou, Li and Burge [10]) is not considered here since it will incur higher delay when the traffic is bursty.

2.3. Combined pickup and delivery

Most work addressing server vacations assume the task is finished once it is served. The scenario we consider, however, requires the message carrier to pick up messages from the foreign host and deliver them to the home host. This is a combined pickup and delivery process. The work that resembles this scenario the most is that of Coffman and Gilbert [11], where they consider a queue and a cart. The tasks served are moved from the queue to a cart, which departs at an appropriate time to deliver all tasks accumulated in it. The main difference of our work is that we aim to provide a *practical* service discipline that works well for *different arrivals* other than just Poisson arrivals, especially bursty ones.

Other work that also considers both pickup and delivery exist for delay tolerant networks [12], [13] and for the vehicular routing problem (VRP) [14], [15]. However, none of these consider the effect of the service discipline, as they usually do not consider the processing time of each task or message.

3. Model Description

As shown in Fig. 1, the pigeon travels at a constant speed, and it takes t_r time to travel from the home host to the foreign host. The message carrier needs to pick up messages from the foreign host through a high-bandwidth wireless local area network (WLAN) interface during each trip. Messages are generated on the foreign host with average rate being $\bar{\lambda}$ and the mean time spent on picking up each message is b. The time needed for the message carrier to transfer messages to the

home host is omitted as it can be in the format of unloading a disk, which takes very limited time.

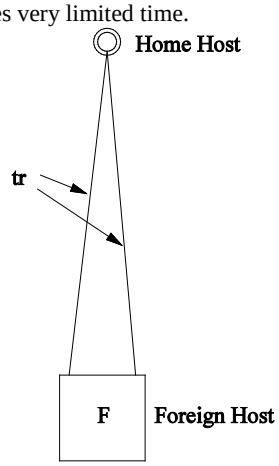

Fig. 1. System model

4. Performance Study Under Non-bursty Traffic

Work on non-bursty arrivals, especially Poisson arrivals, is mostly found in the literature on vacation models. In this section, it is shown that for the scenario studied in this paper, deterministic arrivals have a similar effect as Poisson arrivals, but the analysis is more tractable.

4.1. Exhaustive service discipline

4.1.1. Deterministic arrivals

When an exhaustive service discipline is employed, the message carrier will leave for delivery only when no messages are left in the foreign host. Messages that arrive during the message carrier's pickup process are also picked up. As shown in Fig. 2, a trip begins at the arrival of the message carrier to the foreign host and ends when it returns (from the home host.) Denote the total number of messages picked up on each trip as n_E, and the time spent on pickup as $n_E\bar{b}$. The messages picked up in each trip are generated during the last vacation time and the current pickup time, which is $n_E\bar{s} + 2t_r$. Thus, $\mathrm{n_E} = \bar{\lambda}(\mathrm{n_E}\bar{\mathrm{s}} + 2\mathrm{t_r})$. Solving the equation, we obtain

$$\mathrm{n_E} = \frac{2\bar{\lambda}\mathrm{t_r}}{1-\rho} \qquad (1)$$

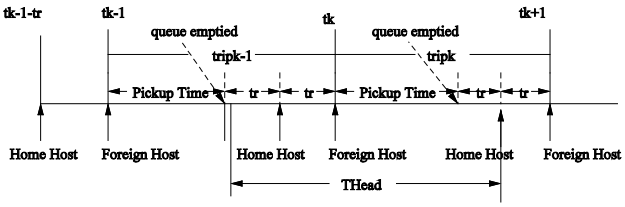

Fig. 2. Dynamic flow of cycles under the exhaustive service discipline

As can be seen from Fig. 2, the time point for the message carrier's departure during *(k −1)*th trip is between the arrival of two messages. With deterministic arrivals it is reasonable to estimate that the first message of the *k*-th trip arrives $\frac{1}{2\bar{\lambda}}$ later. Thus, the delay for the message at the head of each trip is $T_{\square ead} = 2\mathrm{t_r} - \frac{1}{2\bar{\lambda}} + n_E\bar{b} + t_r = n_E\bar{b} + 3t_r - \frac{1}{2\bar{\lambda}}$. The delay of *i*-th message in each trip is $T_{\square ead} - (\mathrm{i}-1)/\bar{\lambda}$. As a result, the average delay for all messages in a trip is

$$\overline{\mathrm{T_E}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n_E}} T_{\square ead} - (\mathrm{i}-1)/\bar{\lambda}}{n_E} = \frac{2-\rho}{1-\rho}\mathrm{t_r} \qquad (2)$$

4.1.2. Poisson arrivals

The work by Coffman and Gilbert [11] about service and delivery using a cart is one of the few analytical papers on the subject of pickup and delivery. Under the assumption of Poisson arrivals, an analysis is given for several different service disciplines. The exhaustive service discipline studied in the vacation model corresponds to the never idle strategy in [11] with no lower threshold on the number of jobs to be picked up ($m = 0$). As presented in [11] (page 877), the average number of messages in the queue-cart system is

$$\bar{\mathrm{n}} = \frac{2\bar{\lambda}\bar{b} + \bar{\lambda}\mathrm{c}^{(2)}/\bar{\mathrm{c}}}{2(1-\bar{\lambda}\bar{b})} + \frac{\bar{\lambda}^2 b^{(2)}}{2(1-\bar{\lambda}\bar{b})^2} \qquad (3)$$

The above formula can be applied to the scenario discussed in this paper. Here b is the time for the pickup of each message, and *c* corresponds to the time that the message carrier is away, which means $c = 2t_r$. The second moments of *b* and *c* are denoted as $b^{(2)}$ and $c^{(2)}$. When the pickup time and the delivery time are both deterministic, $b^{(2)} = \bar{b}^2$, and $c^{(2)} = \bar{c}^2$. Thus, equation (3) becomes

$$\bar{\mathrm{n}} = \frac{\bar{\lambda}\bar{\mathrm{c}}}{2(1-\bar{\lambda}\bar{b})} + \frac{\bar{\lambda}\bar{b}(2-\bar{\lambda}\bar{b})}{2(1-\bar{\lambda}\bar{b})^2} \qquad (4)$$

Applying Little's law, the average delay for each message before the start of delivery is $\frac{\bar{\mathrm{n}}}{\bar{\lambda}} = \frac{\bar{\mathrm{c}}}{2(1-\bar{\lambda}\bar{b})} + \frac{\bar{b}(2-\bar{\lambda}\bar{b})}{2(1-\bar{\lambda}\bar{b})^2}$. Considering the delivery time, which is t_r (or say *c/2*), the average system delay can be obtained as:

$$\overline{\mathrm{T}} = \frac{2-\rho}{1-\rho}\mathrm{t_r} + \frac{\bar{b}(2-\rho)}{2(1-\rho)^2} \qquad (5)$$

Compared to equation (2), which was derived for deterministic arrivals, the only difference is that there is an additional second term: $\frac{\bar{b}(2-\rho)}{2(1-\rho)^2}$. Since the average time for picking up a message $\bar{b}$ is normally much lower compared with the single travel time of a message carrier t_r, this term is negligible only if ρ is not too close to 1. In other words, normally the results obtained for Poisson arrivals and deterministic arrivals do not make any obvious difference. This provides us a foundation for using deterministic arrivals, as it provides for an easier analysis.

If the message lengths are exponentially distributed, the average delay of each message becomes:

$$\overline{\mathrm{T}} = \frac{2-\rho}{1-\rho}\mathrm{t_r} + \frac{\bar{\mathrm{b}}}{2(1-\rho)^2} \qquad (6)$$

which is even closer to the result obtained for the all deterministic distribution case (equation (2)).

4.2. Gated service discipline

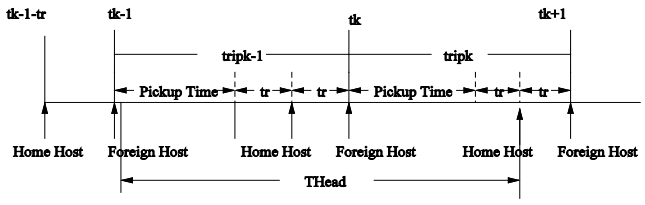

Fig. 3. Dynamic flow of cycles under the gated service discipline

As shown in [10], where a gated service discipline is employed, the relation between the number of messages carried in two continuous trips can be expressed as:

$$n_j = \bar{\lambda}(n_{j-1}\bar{s} + 2t_r) \quad (7)$$

For a stable system, the number of messages picked up at each trip converges to n^*. As shown in [10],

$$n_j = \frac{2\bar{\lambda}t_r}{1-\rho} \quad (8)$$

and the corresponding delay is

$$\overline{T_G} = 2n^*s + 3t_r - \frac{n^*}{2\bar{\lambda}} = \frac{2+\rho}{1-\rho}t_r \quad (9)$$

For the case that the gated service discipline is employed and the arrival is Poisson, an analysis is unavailable for the pickup and delivery scenario. However, simulations using CSIM simulation tools [16] are conducted for comparison. The basic parameters used are: The foreign host is generating a video stream at 1Mbps, and the smallest segment of video must be at least 1 second long so that it is analyzable. This leads to the basic message size being 1M bits. The wireless LAN interface speed is 10Mbps. Thus, the time for picking up a message is b = 0.1. The travel time of the message carrier tr is equal to 600 seconds.

The average delay of each message under different loads is shown in Fig. 4. It can be seen that the results under Poisson arrivals match well with the deterministic arrival case for both exhaustive and gated service disciplines.

4.1.3. Summary

The exhaustive service discipline helps achieve lower average delay compared with the gated service for the non-bursty arrivals. The difference can be as much as 3 times. Actually, this is because the gated service discipline forces messages that arrive before the departure of the message carrier to wait until the next trip. Instead, with the exhaustive service discipline, those messages will be picked up in the current trip. It is also

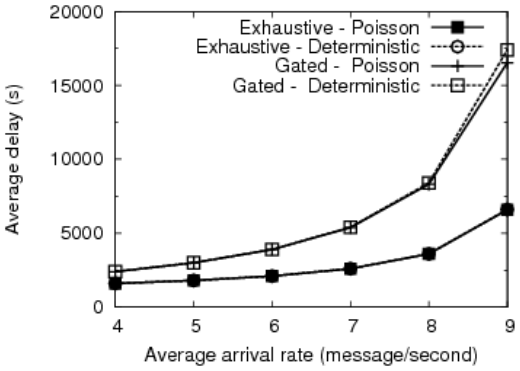

Fig. 4. Comparison of Poisson and deterministic arrival

worth noting that Poisson arrivals do not make much difference from the deterministic arrivals since the pickup process is like batch processing and reduces the variance of the Poisson arrivals to a level that is close to deterministic arrivals.

5. Performance Study Under Bursty Arrivals

5.1. Simplified bursty arrivals

For bursty arrivals, the arrival rates can be extremely high or very low during some periods. The distributions frequently used to describe the bursty traffic include the Modulated Markovian Poisson Process (MMPP) [18], the Correlated Hyper-Exponential Process, and the Interrupted Poisson Process (IPP). All these random processes are composed of two or more phases, with each phase generating Poisson traffic characterized by a different average value. For example, when two phases are used, one phase can be used to represent the peak period with a high arrival rate, and another phase for the valley period with a low arrival rate. For the special case that the arrival rate is equal to 0 in the valley period, it turns into an on/off model which is often modeled by an IPP.

Fig 5. An ideal bursty traffic pattern

To facilitate understanding of the behavior with bursty traffic, we start from the ideally periodic pattern shown in Fig. 5. There are clear peak and valley periods with the arrival rates being λ_{peak} and λ_{valley} respectively. Correspondingly, the length of the peak periods and the valley periods are denoted as t_{peak} and t_{valley}. To make sure the traffic can be dealt with by a message carrier, $\rho = \bar{\lambda}\bar{b} < 1$. However, the arrival rate could be higher than the message carrier's processing capacity during peak periods, which means $\rho_{peak} = \lambda_{peak}\bar{b} > 1$ (the case that $\rho_{peak} < 1$ is not considered in most analysis since it means the arrival is not bursty); while during a valley period, it is surely lower ($\rho_{valley} = \lambda_{valley}\bar{b} < 1$).

As shown by the study in the above Section, an exhaustive service is superior to the gated service for the case that the traffic is relatively steady (e.g. following a deterministic or

Poisson distribution). Will this be still true when the arrival is bursty? In this section we will provide some basic analysis and insights on what kind of service disciplines should be employed when there are obvious peak and valley periods.

For the purpose of comparison, a lower bound of the average delay under the ideal case where the travel time is not considered is analyzed first. Since our main goal is to provide insight, only an approximate analysis for the case when peak traffic dominates is shown, although exact results are provided for the general cases in the numerical results.

5.2. Ideal case - zero-delivery-time

For the most ideal case, we assume that no time is needed for delivery ($t_r = 0$), which means each message that is picked up is delivered immediately. As a result, this is a regular queueing problem with peak/valley arrivals.

The total number of arrivals during the peak period is $n_{peak} = \lambda_{peak} t_{peak}$, and the total time needed to pickup (and deliver) those messages is $\mathbf{n_{peak}}\bar{\boldsymbol{b}}$. The delay of the *k-th* message arriving in the peak period is $\boldsymbol{T_k} = \bar{\boldsymbol{b}} + (\boldsymbol{k}-\mathbf{1})\left(\bar{\boldsymbol{b}} - \frac{\mathbf{1}}{\boldsymbol{\lambda_{peak}}}\right)$, $k = 1...n_{peak}$. For the deterministic arrival case, it can be derived that

$$\overline{T_{peak}} = \frac{T_{n_{peak}} + T_1}{2} = \bar{b} + \frac{(n_{peak}-1)\left(\bar{b} - \frac{1}{\lambda_{peak}}\right)}{2} \approx \frac{(\lambda_{peak}\bar{b} - 1)t_{peak}}{2} = \frac{(\rho_{peak} - 1)t_{peak}}{2} \quad (10)$$

5.3. Exhaustive service

When an exhaustive service discipline is employed, the message carrier starts picking up once it returns, and will continue until the queue becomes empty again. Thus, all messages that arrive during the peak period will be picked up in a single cycle (defined as a *peak cycle*). Although it also includes some of the messages that arrive in the valley period, this part is omitted considering the small number compared with the messages arriving during the peak period.

The delay of the first message that is picked up is approximately $\boldsymbol{T_1} = \boldsymbol{n_{peak}}\bar{\boldsymbol{b}} + \boldsymbol{t_r}$, and for the last message generated at the peak period it is $\boldsymbol{T_{n_{peak}}} = \boldsymbol{n_{peak}}\bar{\boldsymbol{b}} + \boldsymbol{t_r} - \frac{\boldsymbol{n_{peak}}}{\boldsymbol{\lambda_{peak}}}$. Thus, the average delay for the messages in a peak cycle can be estimated as

$$\overline{T_{peak}} = \frac{T_{n_{peak}} + T_1}{2} = n_{peak}\bar{b} + t_r - \frac{n_{peak}}{2\lambda_{peak}} = \rho_{peak}t_{peak} - \frac{t_{peak}}{2} + t_r \quad (11)$$

It can be seen that when the number of messages arriving during the valley period is negligible compared to those that arrive during the peak period, the total delay is about *twice compared with the zero-delivery-time case* (equation (10)).

5.4. Gated service

With a gated service discipline, the number of messages carried in the current cycle are those that are generated during last cycle, thus there is an iterative relationship as shown in equation (7). For the messages arriving during a peak period, denote the number of messages carried in the *j-th* cycle as $\boldsymbol{n_j^{peak}}$ and the total time of this cycle as $\boldsymbol{trip_j^{peak}}$. We have

$$\boldsymbol{n_j^{peak}} = \lambda_{\boldsymbol{peak}}\boldsymbol{trip_{j-1}^{peak}} = \lambda_{\boldsymbol{peak}}\bar{b}\boldsymbol{n_{j-1}^{peak}} + 2\lambda_{\boldsymbol{peak}}t_r . \quad (12)$$

$$\boldsymbol{trip_j^{peak}} = \boldsymbol{n_j^{peak}}\bar{b} + 2t_r = \lambda_{\boldsymbol{peak}}\bar{b}\boldsymbol{trip_{j-1}^{peak}} + 2t_r. \quad (13)$$

The first part in equation (13) is the time spent on the pick-up in each trip, and the second part is the time spent on traveling by the message carrier. The summation of the total *J* trips needed for picking up messages should be equal to the time for picking up n_{peak} messages. Thus,

$$\sum_{j=1}^{J} \lambda_{peak}\, trip_{j-1}^{peak}\bar{b} = n_{peak}\bar{b}. \quad (14)$$

Approximating the relationship between neighboring trips by $trip_j = \rho_{peak} trip_{j-1}$ (recall that $\boldsymbol{\rho_{peak}} = \lambda_{\boldsymbol{peak}}\bar{\boldsymbol{b}}$), after simplification the above equation becomes

$$\frac{\rho_{peak}^{J}-1}{\rho_{peak}-1}\rho_{peak}trip_0 \approx \rho_{peak}t_{peak}, \quad (15)$$

in which $trip_0$ is the time from the start of the peak period to the moment that the message carrier starts to pick up messages. The number of trips needed is

$$J \approx \log_{\rho_{peak}}\left(\left(\rho_{peak}-1\right)\frac{t_{peak}}{trip_0} + 1\right). \quad (16)$$

For the case that $\boldsymbol{\rho_{peak}} \gg \mathbf{1}$, *J* can be approximated as

$$J \approx \log_{\rho_{peak}}\frac{t_{peak}}{trip_0} + 1. \quad (17)$$

Remarks 1:

(1) The number of trips needed is inversely proportional to $trip_0$. In reality, $trip_0$ is usually greater than t_r, which means the message carrier leaves the home host immediately after it knows the demands from the foreign host. If a message carrier waits some time before being deployed, fewer cycles will be needed, and it is more like the exhaustive service case.

(2) To make sure the gated service is more suitable than the exhaustive service, the number of trips needed for carrying the messages in a peak period should be more than 1, which means $J \geq 2$. After checking equation (17) it can be concluded that:

$$t_{peak} \geq \boldsymbol{\rho_{peak}}\, trip_0 \quad (18)$$

The average delay of messages arriving during a peak period under a gated service discipline can be computed as

$$\overline{T_{peak}} = \frac{\sum_{j-1}^{J} n_j^{peak}\overline{T_J^{peak}}}{\sum_{j-1}^{J} n_j^{peak}} = \frac{\sum_{j-1}^{J}\lambda_{peak}trip_{j-1}\left(\frac{trip_{j-1}}{2} + trip_j - t_r\right)}{\sum_{j-1}^{J}\lambda_{peak}trip_{j-1}} . \quad (19)$$

In the above equation, $\overline{\boldsymbol{T_{peak}}} = \frac{\boldsymbol{trip_{j-1}}}{2} + \boldsymbol{trip_j} - \boldsymbol{t_r}$ is the average delay of messages picked up during the *j*-th trip. When the number of peak arrivals and λ_{peak} is high, the travel time t_r is negligible and can be omitted, thus the average delay can be simplified as

$$\overline{T_{peak}} = \frac{\left(\frac{1}{2} + \rho_{peak}\right)\left(1 + \rho_{peak}^{J}\right)}{1 + \rho_{peak}} trip_0 \approx \rho_{peak}^{J} trip_0. \quad (20)$$

Using the expression about $trip_0$ obtained through equation (15), the above equation becomes

$$\overline{T_{peak}} \approx \frac{\rho_{peak}^{J+1}}{\rho_{peak}^{J+1} - 1}\left(\rho_{peak} - 1\right)t_{peak} \approx \left(\rho_{peak} - 1\right)t_{peak}\,. \quad (21)$$

The gated service will usually lead to lower delay than the exhaustive service since there will be several trips instead of one. However, the difference might not be large when λ_{peak} is high such that $\boldsymbol{\rho_{peak}} \gg \mathbf{1}$. In this case the number of cycles (equation (17)) might be small. The delay (equation (21)) is $(\boldsymbol{\rho_{peak}} - \mathbf{1})\boldsymbol{t_{peak}}$, which is very close to the value obtained by the exhaustive service discipline $\boldsymbol{\rho_{peak}t_{peak}} - \frac{\boldsymbol{t_{peak}}}{2}$.

The above analysis reveals the fact that the gated service might not be especially beneficial compared with the exhaustive service. To improve this situation, the number of messages in each peak cycle must be limited, which means the limited-K service discipline might be a better fit.

5.5. Limited-K service

With the limited-K service discipline applied, at most *K* messages can be picked up in each cycle. The number of cycles needed to deliver all messages in the peak period is $C = \frac{n_{peak}}{K}$. Suppose the message carrier arrives before *K* messages are accumulated. The first message in the first group needs to wait $\frac{K}{\lambda_{peak}}$ before the start of a pickup process. Thus, the average delay for the first group is

$$\overline{T_{G,1}} = K\bar{b} + 2t_r + \frac{K}{2\lambda_{peak}}. \quad (22)$$

The delivery time of each group generated during a peak period is $K\bar{b} + 2t_r$ later than the former group. However, it also arrives $\frac{K}{\lambda_{peak}}$ later than the former group. Thus, it can be expressed as

$$\overline{T_{G,J}} = \overline{T_{G,J-1}} + K\bar{b} + 2t_r - \frac{K}{\lambda_{peak}}\,. \quad (23)$$

The average delay for each message can be derived as:

$$\overline{T_{\boldsymbol{peak}}} = \frac{\rho_{peak} - 1}{2}\boldsymbol{t_{peak}} + t_r + \frac{n_{peak}t_r}{K} + K\frac{2 + \rho_{peak}}{2\lambda_{peak}}. \quad (24)$$

The delay will be minimized when

$$K = \sqrt{\frac{2\lambda_{peak}^2 \boldsymbol{t_{peak}} t_r}{2 + \lambda_{peak}\bar{b}}}\,, \quad (25)$$

with the resulted delay being

$$\overline{T_{peak}^{min}} = \frac{\rho_{peak} - 1}{2}\boldsymbol{t_{peak}} + \sqrt{2\left(\rho_{peak} + 2\right)\boldsymbol{t_{peak}}t_r} + t_r = \left(\sqrt{\frac{\left(2 + \rho_{peak}\right)t_{peak}}{2}} + \sqrt{t_r}\right)^2 - \frac{3t_{peak}}{2}, \quad (26)$$

If $\sqrt{\frac{(2+\rho_{peak})\boldsymbol{t_{peak}}}{2}} \gg \sqrt{t_r}$, which means the time spent on picking up messages arriving during a peak period ($\rho_{peak}\boldsymbol{t_{peak}} = \lambda_{peak}\boldsymbol{t_{peak}}\bar{b} = n_{peak}\bar{b}$) is much higher than the round trip time of the message carrier ($2t_r$), the average delay of each message under this limited-K service discipline is close to the zero-delivery-time case ($\frac{\rho_{peak}-1}{2}\boldsymbol{t_{peak}} + t_r$).

5.6. Comparison of service disciplines under bursty arrivals

The average delay of each message under different service disciplines is shown through numerical examples as in Fig. 6. The basic parameters used are the same as in Section IV: $\bar{b} = 0.1, and\ t_r = 600s$. The average arrival rate is 5 Mbps, which means $\bar{\lambda} = 5$ messages/sec. For simplicity, we assume that only one parameter is needed to determine the rate and length relations between the peak and valley period. That is, assume $\lambda_{peak} = a\bar{\lambda}, \lambda_{valley} = \frac{\bar{\lambda}}{a}, t_{valley} = at_{peak.}$ This relation guarantees that the average arrival rate in the whole period is $\bar{\lambda}$ irrespective of the change of parameter *a*.

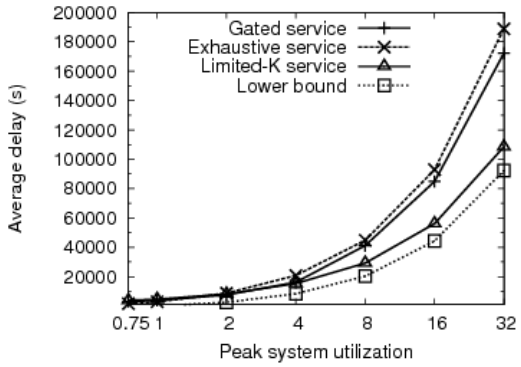

Fig. 6. Comparison under different peak loads (peak period = 6000s).

In Fig. 6, the length of the peak period is fixed at 6000 seconds. With the change of the system utilization at the peak period ρ_{peak}, the average delay also varies. With high values of ρ_{peak}, the gated service performs better than the exhaustive service. However, the benefit is not as obvious as the limited-K service, especially when the traffic is very bursty. Actually it can be observed that the average delay of each message under the limited-K service discipline is close to the zero delivery- time case, which is about half compared to the delay under the exhaustive service discipline.

The above numerical results verify our analysis about different service disciplines. Although the above conclusion is made based on the study of periodic peak/valley arrivals, it can help provide heuristics for dealing with random arrivals with bursty characteristics.

6. Dynamic Service Disciplines For General Arrivals

Although Coffman and Gilbert [11] give analysis on exhaustive and different possible limited service disciplines, they could not provide a way to find the optimal service discipline.

Actually, they resort to simulations for searching a possibly optimal one. This means that the optimal service discipline in terms of achieving minimum average delay is hard to obtain, even for Poisson arrivals. In fact, the optimal service discipline is very sensitive to the parameters of the arrival distribution, which makes it more practical to find heuristic methods that can adapt well to real traffic with general arrivals. The resulting delay might not be the minimum, but should be close to optimal for most cases.

There are at least two justifications for proposing an adaptive service discipline:
(1) The messages are generated dynamically in the foreign host, and it is usually unrealistic for the message carrier to know the arrival distribution in advance. The service discipline should be flexible enough to deal well with both bursty and non-bursty arrivals;
(2) Even for a bursty arrival pattern, there are periods that are non-bursty (valley periods) and it might be beneficial to use different service disciplines at peak and valley periods.

6.1. Adaptive service discipline

For the periodic arrivals studied above, a peak period is followed by a clearly separated valley period. The valley period is also long enough that it will not be combined into the next peak period. This clear separation is convenient for analysis as only a period need to be analyzed. When arrivals are random, some valley periods might be short, so they will be incorporated into a larger peak period by the message carrier. Due to the ambiguity of the border between the peak and valley periods, similar analytical methods become extremely difficult and we need to resort to heuristic methods.

The heuristic methods we propose are based on the analysis presented in Section 4 and Section 5: the gated or limited-K service discipline should be used for dealing with bursty periods, and the exhaustive service has the advantage during non-bursty periods. Thus, a major task is to determine whether the arrivals occur during a bursty or non-bursty period, which can be only be judged by limited knowledge acquired through real-time measurement.

To avoid frequent switching of service disciplines, the long term trend should play a more important role in judging the peak/valley periods. To this end, the change of the queue length upon the return of the message carrier is used as the judgment on the trend of change of the arrival process. When the arrival process is in its peak period, the messages will accumulate in the queue. This increasing accumulation can be detected by the message carrier with basic measuring abilities. For example, it can be detected by measuring the number of messages in the queue seen by a returning message carrier. If the number of messages in the queue upon a message carrier's arrival is more than what was delivered in the last trip, it is an indication of a peak period, otherwise it indicates a valley period.

A problem with this simple criteria is that it might be sensitive to the change of arrival rate. For instance, when $n_k < n_{k-1}$, an exhaustive service will be employed. If a burst arrives before the k_{th} trip departs, the delay might be extremely high. To prevent this, an additional constraint can be added to ensure that the number of messages picked up using the exhaustive service will not surpass that of the former trip.

The adaptive service discipline algorithm can be summarized as:

Algorithm 1: Adaptive gated/exhaustive service discipline
1. A message carrier measures the number of messages in the queue upon its return to the foreign host and the number of messages to be delivered upon its departure.
2. If the number of messages in the queue is more than what was delivered in the last trip and the estimated peak period is longer than the initial trip time (equation (18)), a gated service discipline is employed; otherwise an exhaustive service is taken.
3. When an exhaustive service is employed, the message carrier must depart when the queue is empty or has accumulated more messages than the last trip.

6.2. Performance of the adaptive service discipline on random arrivals

In scenarios such as a foreign host caused by an earthquake, the generation of messages, which are merged from different sensors distributively located, generally have strong spatial and temporal correlations. This kind of correlation is usually shown through bursty arrivals that last for some time. For instance, an additional small quake will trigger the sensors and monitors deployed in the affected area resulting in the generation of high speed data for a certain time (say half hour). Then most sensors go back to sleep, and the traffic rate will remain low until sensors are triggered by a new event.

The application of the adaptive service discipline in such typical scenarios is shown in the following example.

The arrival process is composed of peak segments and valley segments, where each segment has the same number of messages, but generated at peak or valley rates. The pseudo code for generating the messages is as follows:

```
for(; ; )
if (prob() < 1/11)then
   for(i = 0; i < segment length; i++)
      hold(expntl(peak interarrival time))
else
   for(i = 0; i < segment length; i++)
      hold(expntl(valley interarrival time))
```

From the code it can be seen that 1/11 of total messages are generated at the valley period, and 10/11 of them are generated at the peak rate. The peak rate is 10 times of the average rate, and the valley rate is 1/10 of the average rate, which means the traffic mode is similar to what we used in Section V by setting a = 10. Correspondingly, the total length of the valley period is 10 times that of the peak period. The average arrival rate is $\bar{\lambda} = \mathbf{7}$, and all other parameters are same as in Section IV.

To show the effectiveness of our adaptive service discipline, it is compared with both gated and exhaustive service disciplines through simulations as in Fig. 7.

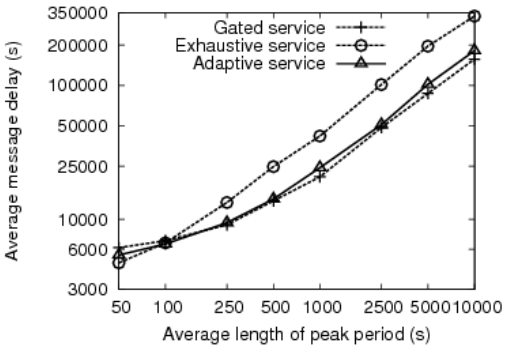

Fig. 7 Comparison of different service disciplines.

It can be seen that when the average duration of the peak arrival period is short (e.g. 50 seconds), the exhaustive service works the best. This is because the number of messages seen by a returned message carrier is relatively low and steady, which is similar to Poisson arrivals.

When the duration of each peak arrival period is long (e.g. 2500 seconds), the length of a single peak period goes up dramatically and makes the gated service much more favorable than the exhaustive service. The adaptive service discipline, although possibly not optimal, behaves always close to the optimal one. Thus, it is a good choice for the message carrier since the traffic rate and pattern are not known ahead of time.

6.3. Adaptive limited service

As shown above, the adaptive service using a combination of gated and exhaustive service can adapt to the change of peak and valley periods well. However, the gated service is not the best service for dealing with long bursty arrivals. Thus, an improved algorithm that uses the limited-K service in place of the gated service in the adaptive service discipline is shown below.

Algorithm 2: Adaptive limited/exhaustive service discipline

1. A message carrier measures the number of messages in the queue upon its return to the foreign host and the number of messages to be delivered upon its departure.
2. If the number of messages in the queue is more than what was delivered in the last trip and the estimated peak period is longer than the initial trip time, a limited-K service discipline is employed; otherwise an exhaustive service is used.
3. To compute the optimal K value, the average arrival rate 8 during the last trip time is estimated: the number of new arrivals during the last cycle time divided by the length of the last cycle time. The optimal group size K is determined using equation (25). The length of peak period is estimated as the summation of continuous peak cycle times.
4. When the queue size decreases during the pickup process, the limited-service continues until the queue size is reduced to a low level (e.g. $< \frac{2t_r}{\bar{b}}$).
5. When an exhaustive service is used, the message carrier must depart either when the queue is empty or when the messages gathered is more than in last trip.

Note that in this algorithm the value of K might change in each cycle since the estimated peak rate and peak period of each cycle might vary. When the queue size decreases during the pickup process, the peak period is assumed to be over, but the limited-K service still continues (till the queue size drops to a certain low level) with the estimated peak period unchanged. The comparison of this adaptive limited service discipline with the gated and the adaptive gated/exhaustive service is shown in Fig. 8. Obviously, when the bursts last long, the adaptive limited service discipline can achieve even lower delay than the gated service discipline.

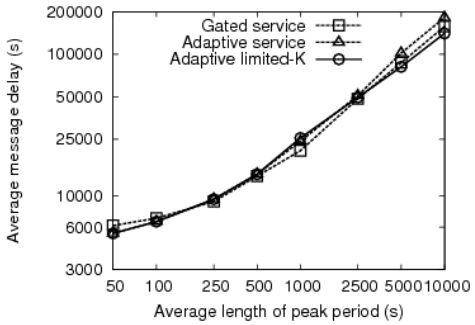

Fig. 8. Comparison of adaptive limited service discipline

It is also worth noting that the estimations of the peak period and peak rate under a random arrival pattern are generally not that accurate, which makes the benefits of using the limited- K service not as obvious compared with a more deterministic arrival pattern used in Section 5.

7. Conclusion

We studied the scheduling of message carriers for message pickup and delivery in a pigeon network. When arrivals are non-bursty, the exhaustive service discipline is optimal. For the bursty case, analysis is provided for a basic periodic peak/valley arrival process, and we conclude that the gated and limited-K service disciplines are superior to the exhaustive service discipline.

As a practical implementation, this paper introduces adaptive service disciplines that can change the service discipline according to the arrival process. The numerical results show that the proposed algorithms perform well under different traffic situations. This demonstrates that this kind of dynamic service discipline is a good option for scheduling the message carrier, especially when there is no *a priori* knowledge about the arrival process.

For future work, optimal scheduling strategies can be explored when the deadlines of messages are considered or when multiple message carriers are available.

References

[1] K. Fall, “A delay-tolerant network architecture for challenged internets”, SIGCOMM 2003

[2] M. Eisenberg, “Queues with periodic service and changeover times”, Operation Researches, Vol.20, pp.440–451, 1972 http.dx.doi.org/10.1287/opre.20.2.440

[3] Y. Levy and U. Yechiali, “Utilization of Idle Time in an M/G/1 Queueing System”, Management Science, Vol. 22, No.2, October 1975

[4] S. Fuhrmann and R. Cooper, “Stochastic Decompositions in the M/G/1 Queue with Generalized Vacations”,

Operations Research, Vol. 33, No. 5 (Sep. - Oct., 1985), pp. 1117-1129 http.dx.doi.org/10.1287/opre.33.5.1117

[5] B. Doshi, "A Note on Stochastic Decomposition in a GI/G/1 Queue with Vacations or Set-up Times", Journal of Applied Probability, Vol. 22, No. 2 (Jun., 1985), pp. 419-428 http.dx.doi.org/10.2307/3213784

[6] M. Yadin and P. Naor, "Queueing Systems with a Removable Server", Operational Res. Quart. Vol. 14, pp.393–405,1963. http.dx.doi.org/10.1057/jors.1963.63

[7] Thomas B. Crabill, Donald Gross, Michael J. Magazine, "A Classified Bibliography of Research on Optimal Design and Control of Queues", Operations Research, Vol. 25, No. 2 (Mar. - Apr., 1977), pp. 219–232 http.dx.doi.org/10.1287/opre.25.2.219

[8] L. Tadj and G. Choudhury, "Optimal Design and Control of Queues", Sociedad de Estadistica e Investigacin Operutiva Top (2005) Vol. 13, No. 2, pp. 359–412

[9] J. Moder and C. Phillips, "Queuing with Fixed and Variable Channels", Operations Research, Vol. 10, No. 2 (Mar. - Apr., 1962), pp. 218–231 http.dx.doi.org/10.1287/opre.10.2.218

[10] J. Zhou, J. Li, and L. Burge III, "Efficient scheduling of pigeons", Eurasip Journal on Wireless Communication and Networks no.3, March 2010.

[11] E. Coffman and E. Gilbert, "Service by a queue and a cart", Management Science, vol.38, no.6, June 1992.

[12] W. Zhao and M. Ammar, "Message Ferrying: Proactive Routing in Highly-Partitioned Wireless Ad Hoc Networks," Proceedings of the IEEE Workshop on Futrure Trends in Distributed Computing Systems, May 2003.

[13] W. Zhao, M. Ammar, and E. Zegura, "Controlling the Mobility of Multiple Data Transport Ferries in a Delay-Tolerant Network," IEEE INFOCOM 2005.

[14] M. Savelsbergh and M. Sol, "The general pickup and delivery problem", Transportation Science, Vol.29, pp.17–29, 1995. http.dx.doi.org/10.1287/trsc.29.1.17

[15] M. Swihart and J. Papastavrou , "A stochastic and dynamic model for the single-vehicle pick-up and delivery problem". European Journal of Operational Research, Vol. 114, No. 3. (01 May 1999), pp. 447–464. http.dx.doi.org/10.1016/S0377-2217(98)00260-4

[16] Mesquite Software, Inc. "CSIM19 User's Guide", Austin, Texas, 2001.

[17] D. Lucatoni, K. Meier-Hellstern and M. Neuts, "A single server queue with server vacations and a class of non-renewal arrival processes", Advanced Applied Probability, Vol.22, pp.676-705, 1990 http.dx.doi.org/10.2307/1427464

[18] H. Heffes and D. Lucatoni, "A Markov Modulated Characterization of Packetized Voice and Data Traffic and Related Statistical Multiplexer Performance", IEEE Journal on Selected Areas in Communications, Vol. SAC-4, No. 6, September 1986