Quantitative Analysis of Cloud-based Streaming Services

Fang Yu[1], Yat-Wah Wan[2] and Rua-Huan Tsaih[1]

1. Department of Management Information Systems
National Chengchi University, Taipei, Taiwan
2. Graduate Institute of Logistics Management
National Dong Hwa University, Hualien, Taiwan

June 14, 2013

IEEE SCC 2013

This work has been accepted to be published in:
10th IEEE International Conference on Service Computing
Santa Clara, June 2013.

Cloud-based Services

Services deployed on *Clouds* are getting more and more popular

- Commerce: online banking, online shopping, etc.
- Entertainment: online music and videos, gaming, etc.
- Interaction: social networks, blogger, etc.

Advantages

- Service Scalability and Availability
- Dynamic Resource Allocation

Servie Quality

Managing high-standard service quality becomes a critical issue for online business success

As for the National Palace Museum project, we are particularly interested in online cloud-based streaming services.

Theoretical Gap

Intuitively, the service quality is related with:

- the processing ability on the server side,
- the bandwidth allocated for the customer,
- the traffic condition of the Internet, and
- the processing ability of the end device on the client side.

There is a lack of formal quantitative analysis, nor theoretical exploration, of the relationship between the service quality and these four factors, particularly for cloud-based services.

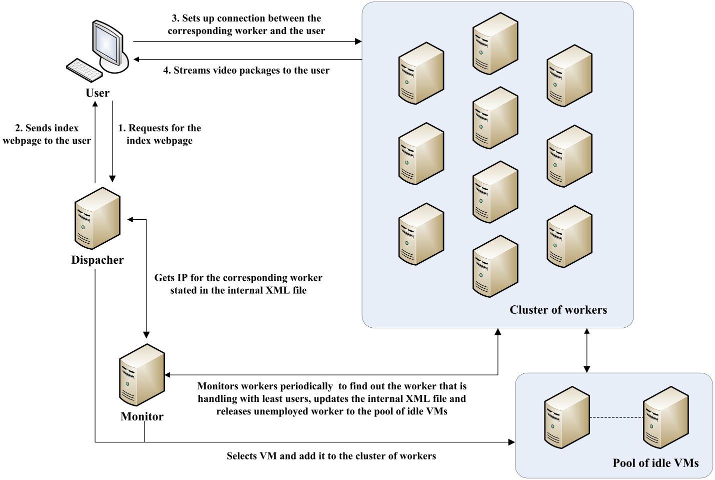
Objectives

This study addresses this theoretical gap with:

- formalizing the service framework for cloud-based streaming services with queuing models
- proposing macro models for deriving operations characteristics of systems with closed form expressions (from the system perspective)
- proposing micro models with simulation procedures for observing service dynamics and indicators (from the customer perspective)

Cloud-based Service System

Using iPalace Video Channel as an example:


3. Sets up connection between the corresponding worker and the user
4. Streams video packages to the user
User
2. Sends index webpage to the user
1. Requests for the index webpage
Dispacher
Gets IP for the corresponding worker stated in the internal XML file
Cluster of workers
Monitor
Monitors workers periodically to find out the worker that is handling with least users, updates the internal XML file and releases unemployed worker to the pool of idle VMs
Selects VM and add it to the cluster of workers
Pool of idle VMs

The Two-phase Service Framework

- PHASE I:
  1. The arrival comes to the *Dispatcher*.
  2. The *Dispatcher* responds with the corresponding worker.
- PHASE II:
  1. The arrival is redirected to the corresponding worker.
  2. The corresponding worker provides the subsequent video service (similar to a television program) via broadcasting the selected video streaming that interweaves with the advertisements until the arrival reneges.

The Queuing Model of Phase I

The service system regarding the phase I is a single-server service system that every arrival is served by the *Dispatcher*.

Phase I is an $M/D/1 : \infty/\infty/FCFS$ queuing system:

1. Arrivals are served on a FCFS (first come, first serve) basis.
2. Arrivals are independent of preceding arrivals, but the arrival rate λ does not change over time.
3. Arrivals are described by a Poisson probability distribution and come from a very large population.
4. Service times for all arrivals are fixed and known (a constant denoted as d).
5. d is negligible, compared with the value of $\frac{1}{\lambda}$.

The Queuing Model of Phase I

In this $M/D/1 : \infty/\infty/FCFS$ queuing system, the service time is fixed as d (to have a stable system, λd needs to be smaller than 1).

- the stationary-state probability of zero arrivals in the system P_0 equals $1 - \lambda d$
- the average number of customers in the service system L equals $\frac{2\lambda d - \lambda^2 d^2}{2(1-\lambda d)}$,
- the average number of arrivals waiting in queues L_q equals $\frac{\lambda^2 d^2}{2(1-\lambda d)}$,
- the average time spent in the service system W equals $\frac{2d - \lambda d^2}{2(1-\lambda d)}$, and
- the average time spent waiting in queues W_q equals $\frac{\lambda d^2}{2(1-\lambda d)}$.

The Queuing Model of Phase II

- Arrival: With the assumptions of the Phase I system, the departure process of Phase I is the same as the arrival process of Phase II. Thus the arrival process of Phase II is a Poisson process with mean rate λ.
- Departure: It is the reneging behavior of the customer that defines the service time of the customer with the corresponding worker. Such service times are assumed to be independent and identically distributed across customers with a known distribution.

The Queuing Model of Phase II

The service system in phase II is an $M/M/1 : \infty/\texttt{BOUND}/\textit{round-robin}$ queuing system.

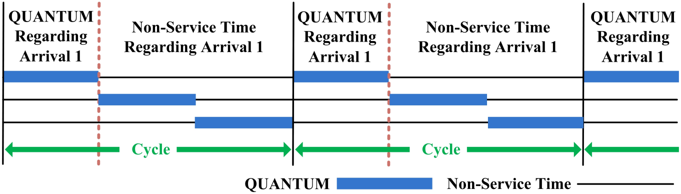
- Arrivals are served on a *round-robin* basis.
- Arrivals follow the Poisson process of rate λ, independent of the service process.
- The time `WaVidtime` for a customer spent on watching video are i.i.d. negative exponential probability distribution of rate μ, a known positive constant. Thus, P($\texttt{WaVidtime} \leq x$) $=$ 1 - $e^{-\mu x}$ for $x \geq 0$.

The Queuing Model of Phase II

- Assuming the initial setup and communication time are relative small and negligible, it follows that the service time for a worker spent on an arrival (i.e., the duration from the epoch that the worker starts to generate video packets for the customer to the epoch that the worker stops to generate video packets for the customer) are also i.i.d. negative exponential probability distribution of rate μ.
- Each worker serves at most the BOUND amount of arrivals. Once the value of LENGTH hits the BOUND, no new customer will be assigned to this worker, which will retire from service after clearing the customers on hand.

Round-Robin Schedule

The worker adopts the *round-robin* algorithm (Silberschatz et al. 2004) to handle tasks of all arrivals and spends QUANTUM in each arrival for generating the packages of video, while the remaining tasks of the other arrivals are waiting in queue without being executed.


QUANTUM Regarding Arrival 1
Non-Service Time Regarding Arrival 1
QUANTUM Regarding Arrival 1
Non-Service Time Regarding Arrival 1
QUANTUM Regarding Arrival 1
Cycle
Cycle
QUANTUM
Non-Service Time

Quantitative Analysis

The key questions for the manager of online streaming services to answer are:

- the operations characteristics of the system, which can be measured by indicators such as the average number of customers in system and the average number of VMs used, and
- the service quality for customers such as experienced lag time and interruptions

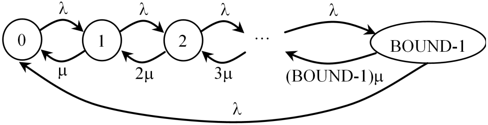
Macroscopic Model: Model for the Number of Customers in System

Let $s = (n;\, n_1, \ldots, n_{\mathtt{BOUND}})$ be the state of the system, where n is the number of customers at the corresponding worker, $n \le \mathtt{BOUND} - 1$. n_j is the number of retiring workers with j customer at the machine, $j \in \{1..\mathtt{BOUND}\}$.

- Total rate out: $q_s = \left[\lambda + \left(n + \sum_{j=1}^{\mathtt{BOUND}} j \times n_j\right) \times \mu\right]$;
- Effect of an arrival:
  - $q_{s,s_+} = \lambda$, for $n \le$ BOUND-2;
  - $q_{s,R} = \lambda$, for $n =$ BOUND-1;
- Effect of a departure:
  - $q_{s,s_-} = n \times \mu$
  - $q_{s,s_{j-}} = n_j \times \mu \times j$, for $n_j \ge 1$, $\forall j \in \{1..\mathtt{BOUND}\}$.

Stationary Distribution of the State of the Corresponding Worker

We derive the stationary distribution of the state of the corresponding worker when a new arrival comes in. The state is defined as the number of customers served by the corresponding worker (before the new arrival).



Stationary Analysisl

When stable, we have:

$$\begin{aligned}
\lambda \times p_{\mathtt{BOUND}-1} + \mu \times p_1 = \lambda \times p_0,\\
\lambda \times p_{i-1} + (i+1) \times \mu \times p_{i+1} = \\
(\lambda + i \times \mu) \times p_i, \forall i \in \{1..\mathtt{BOUND}-2\},\\
\lambda \times p_{\mathtt{BOUND}-2} = (\lambda + \mu \times (\mathtt{BOUND}-1)) \times p_{\mathtt{BOUND}-1}.
\end{aligned} \tag{1}$$

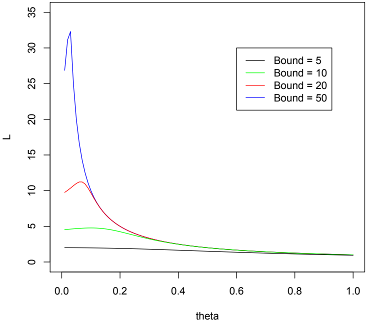
Stationary Analysis

Applying the normalization equation to Eqt. (**??**), we have, for $\kappa \in \{1..\mathtt{BOUND}\}$, the closed form solution:

$$p_{\mathtt{BOUND}-\kappa} = \frac{\sum_{m=1}^{\kappa} \left(\prod_{j=m}^{\kappa-1}(\mathtt{BOUND}-j)\right) \times \theta^{k-m}}{1 + \sum_{k=2}^{\mathtt{BOUND}} \left[\sum_{m=1}^{k} \left(\prod_{j=m}^{k-1}(\mathtt{BOUND}-j)\right) \times \theta^{k-m}\right]}. \qquad (2)$$

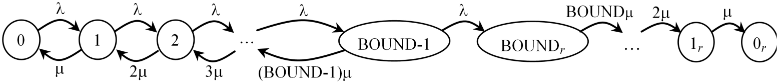
Average Number of Customers (Seen by a New Arrival)

$$L = \sum_{k=0}^{\mathtt{BOUND}-1} p_k \times k, \tag{3}$$



The Mean Time for a Virtual Machine from Activated to Idle

We deduce the mean time taken for a worker to go through the process of being activated, followed by retirement, and then eventually becomes unemployed (to the idle pool).



The Mean Time for a Virtual Machine from Activated to Idle

Within one round of duty, on average a worker works for E($T_{0,0_r}$) $= \mathrm{E}(T_{0,\mathtt{BOUND}_r}) + \mathrm{E}(T_{\mathtt{BOUND}_r,0_r})$ units of time, and on average accepts $\lambda\times$ E($T_{0,\mathtt{BOUND}_r}$) arrivals. Because the rate of customers handled by one worker is

$$\frac{\lambda \times E(T_{0,\mathtt{BOUND}_r})}{E(T_{0,\mathtt{BOUND}_r}) + E(T_{\mathtt{BOUND}_r,0_r})},$$

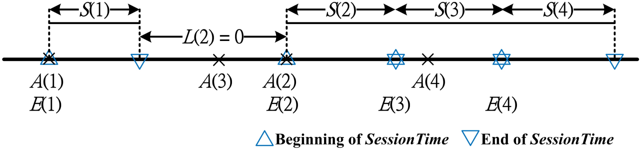
on average there are

$$1 + \frac{E(T_{\mathtt{BOUND}_r,0_r})}{E(T_{0,\mathtt{BOUND}_r})}$$

workers employed by the system.

Microscopic Model: Model for the Service Experience of Customers

Given the time-sharing dynamics at `QUANTUM` level of workers, the traffic condition of WWW, and the properties of the end device of the customer, the microscopic model discusses procedures that eventually give the (expected) amount of lag time and the number of interruptions experienced by a customer.



Arrival, Effect and Lag Time of the ith Packets

$$
\begin{aligned}
&ArrivalTime(i) = InitServer + \\
&\sum_{j=1}^{i-1} (\mathtt{LENGTH}_j \times \mathtt{QUANTUM}) + \mathtt{QUANTUM} + \\
&NormalTransTime(i) + DelayTransTime(i), \forall i \geq 1; \qquad (4)
\end{aligned}
$$

$$
\begin{aligned}
EffectTime(i) =& \mathtt{max}\{EffectTime(i-1) + \\
&SessionTime, ArrivalTime(i)\}, \forall i \geq 2. \qquad (5)
\end{aligned}
$$

$$
\begin{aligned}
LagTime(i) = [&ArrivalTime(i) - (EffectTime(i-1) + \\
&SessionTime)]^{+}, \forall i \geq 2. \qquad (6)
\end{aligned}
$$

Simulation

Simulation procedure for the End Device Queue for a Customer Arriving at state k with $k < \texttt{BOUND}$ with given $\texttt{QUANTUM}$, $SessionTime$, $NormalTransTime()$, $DelayTransTime()$.

parameter	definition
NI	the cumulative number of interruptions within the $WaVidTime$ of a customer
LT	the corresponding cumulative lag time
NR	the number of replications in the simulation procedure
p_u	the probability of the state going up
I_k	the estimate of expected number of interruptions in $WaVidTime$ for a customer who on arrival finds k customers with the corresponding worker
L_k	the estimate of the expected duration of lag time for a customer who on arrival finds k customers
s	the current number of customers i
rb	a flag for being in the retiring stage

Simulation

We simulate NR times of a sample term. Each sample term simulates the process of a customer from its arrival to its departure, collecting accumulated lag time and interruptions that the customer experiences.

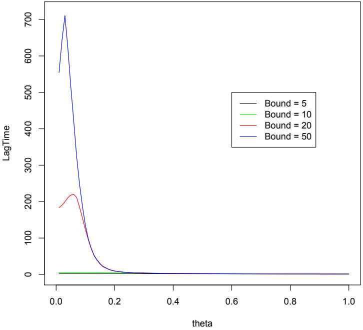
1. Set $NI = 0$, $LT = 0$, $t_0 = 0$, $m = 1$. Select a value for NR.
2. While $(m \leq NR)$ simulate each term
3. For a customer arriving at state k, $I_k = NI/NR$, and $L_k = LT/NR$.

Simulation

1. Set n=1, s=k+1, rb=`False`;
2. If rb=`True` or s = `BOUND`, set $\eta = s \times \mu$, p_u = 0, and rb=`True`; else $\eta = \lambda + s \times \mu$ and $p_u = \frac{\lambda}{\lambda + s\times\mu}$.
3. Draw a random variate t_n-t_{n-1} from $\sim$exp(η). The worker stays at state s in (t_{n-1}, t_n).
4. Simulate the end device queue according to Eqt. (4), (5) and (6) with `LENGTH` = s. Set NI = NI+1 whenever an interruption occurs, and accumulate the corresponding lag time of the customer in (t_{n-1}, t_n) to LT.
5. At t_n, set δ = 1 with probability p_u, and δ = -1 with probability (1-p_u). If δ = -1, go to 2.6 with probability $\frac{1}{s}$ else set n = n+1, s = s+δ, and go to 2.2.
6. Set m = m+1.

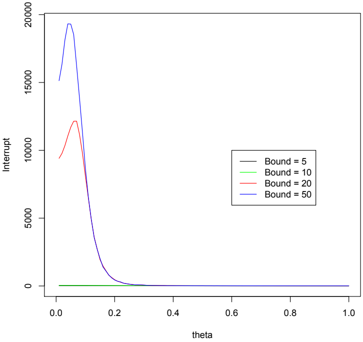
Lagtime

$$Lagtime = \sum_{k=0}^{\mathtt{BOUND}-1} p_k \times L_k, \tag{7}$$



Interrupt

$$Interruption = \sum_{k=0}^{\mathtt{BOUND}-1} p_k \times I_k, \tag{8}$$



Conclusion

As online streaming services significantly increase in recent years, it becomes a critical issue to offer quality service via systems that benefit from cloud computing developments.

We pioneer the study on quantitative analysis to estimate and further improve the quality of cloud-based streaming services, deriving theoretical results on operations characteristics of queuing models with mild assumptions.

By simulating the continuous-time Markov chain according to the adopted operations rules for VMs, we also get performance indicators such as the average lag time and interruptions that a customer may experience under different environment settings.

Ongoing work

The presented approach provides managers of online streaming services a formal and systematic way to evaluate service quality before launch. One of our ongoing work is applying the presented approach to analyze iPalace Channel (Tsaih et al. 2012), a digital exhibition channel of National Palace Museum (NPM) in Taiwan.

Any suggestion?

Thank you for your attention.