Decision Support Systems 77 (2015) 55–66

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Resource allocation and revenue optimization for cloud service providers

Jhih-Hua Jhang-Li [a], I. Robert Chiang [b,*]

[a] Department of Information Management, Hsing Wu University, Taiwan
[b] Gabelli School of Business, Fordham University, United States

ARTICLE INFO

Article history:
Received 21 March 2014
Received in revised form 27 February 2015
Accepted 13 April 2015
Available online 1 May 2015

Keywords:
Versioning
Cloud service provider
Priority queues
Advertising
Personalization

ABSTRACT

Online storage and streaming services are surpassing physical media as the predominate means of disseminating and sharing digital contents such as music, documents, photos, and videos. In addition, many software vendors are switching from on-premises installations to web-based rendering for their offerings. Differentiated pricing, based on tiered service responsiveness and advertisement displays, has been widely adopted by cloud service providers to optimize resource utilization and improve profitability under heterogeneous user demands. We analyze the impact of resource allocation and advertising decisions on provider profit and social welfare when separating premium subscription from more basic offerings. By considering queuing delays and advertising revenues, we suggest conditions under which the service provider should invest in service quality to grow the subscription base, which in turn helps attract more advertisers. We also assess the impact of advertising technology that lessens the users' disutility toward advertisements and increases the likelihood of ads click-through. Finally, we point out when offering free services could be more profitable than charging a subscription fee.

1. Introduction

The latest wave of technology advancement has made cloud-based distribution the primary channel for digital content (e.g., documents, mails, videos, images, music, and gaming). Compared with on-premises installations, cloud-based distribution enables content, application, and service providers to maintain better customer relationships (e.g., via analyzing usage patterns), faster updates/patches, and more efficient delivery of marketing offerings. Online service delivery has shown to be a viable and profitable business model. For example, a file-sharing site brought in $175 million of profits through premium membership fees and sponsored advertisements while accounting for a significant percentage of file sharing traffic.[1] Worldwide, the market for cloud-based content distribution is projected to reach $4bn by 2016.[2]

An important issue when managing the cloud delivery platform is to balance the revenue streams from subscription and advertising. When managing user subscriptions, quality discrimination by offering differentiated features and service levels is a common approach that allows users to self-select the pricing tier per their willingness to pay. For example, one major web mail service is free, provided that the email contents are displayed with sidebar ads; the Plus version of the same service charges an annual subscription but is ads-free. While over-exposure to ads may be a turn-off to some users, content and service providers have come to rely on sponsored advertisements as a steady source of revenue.[3]

Congestion delay is a key measurement for service quality, thus many design considerations for cloud storage and service platforms are related to (downloading and uploading) throughput rates [44]. Due to limits in network bandwidth, storage capacity, server speed, security measures, and other infrastructure bottlenecks, improving the service quality in terms of site responsiveness likely requires significant investment in the service platform. The perceived responsiveness can be further regulated by scheduling policies such as First-Come-First-Served (FCFS) and priority scheduling. In FCFS, all requests are processed according to their order of arrival. With priority scheduling, requests within each service class are queued as they arrive; requests from higher-priority accounts are moved ahead of those from lower-tier clients. Empirically, some service providers[4] treat requests from their paid customers with priority, while others[5] apply FCFS policy to all requests. The design of service quality and pricing levels has far-reaching impacts. For example, it has been observed that five percent

* Corresponding author at: Department of Information Management, Hsing Wu University, No. 101, Sec. 1, Fenliao Rd., LinKou District, New Taipei City 244, Taiwan. Tel.: +886 2 26015310x2461.
E-mail addresses: jhangli@gmail.com (J.-H. Jhang-Li), ichiang@fordham.edu (I.R. Chiang).

[1] http://www.digitaltrends.com/web/megaupload-effect-filesonic-drops-file-sharing-uploaded-to-drops-the-us/.
[2] http://www.statista.com/statistics/203474/global-forecast-of-cloud-computing-storage-services-revenue/.
[3] http://www.iab.net/media/file/IAB_Internet_Advertising_Revenue_Report_FY_2013.pdf.
[4] For example, FileFactory (http://www.filefactory.com/) and Rapidshare (https://rapidshare.com/).
[5] For example, FileHippo (http://www.filehippo.com) and Ziddu (http://www.ziddu.com/).

http://dx.doi.org/10.1016/j.dss.2015.04.008
0167-9236/

of a major telco's users generate roughly half of the data traffic[6]; a file hosting service provider reported that roughly 90% of file downloads were made by its free users [28].

Capacity allocation and planning have been studied in a wide variety of queuing optimization situations for firms to expand or reduce service rates as demands fluctuate; there have also been voluminous reports on advertisement versioning[7] and personalization.[8] However, few studies thus far have linked capacity planning and service versioning even though such a two-pronged approach is widely implemented by service providers. For instance, an anonymous user[5] receives content at throttled speed with popup and banner ads during file transfer; a paid user enjoys a service with higher download rate without ads.

Combined, user segmentation, service pricing, and platform investment decisions closely link to the service firm's profitability. This study aims to show how a cloud service provider can best segment its service along the dimensions of data rates and advertising levels by answering two key research questions: (1) How does a cloud service provider adjust its service capacity and advertising level as advertising technology advances? (2) What is the impact of advertising technology on subscription fees under different scheduling policies (i.e., FCFS vs. priority) and advertising programs (i.e., generic vs. stratified/customized ads display)?

2. Literature review

From a more technical perspective, digital content can be distributed through IP networks using either client–server processing or peer-to-peer (P2P) technology. Despite P2P's wide popularity [14], the client–server model remains essential because P2P download speeds are more erratic for depending on the number of peer nodes involved and the peer nodes' allotted upload bandwidth. HTTP-based client–server media sharing, on the other hand, is popular when serving premium users because preferred digital content can be rendered at a sustained high speed [35].

2.1. Advertising

Advertising is a business based on creating attention. Online providers draw attentions with its digital content and applications, then monetize the attention in form of advertising revenue [13]. The “fremium” model with ads-supported basic offering is very common among mobile apps, email services, and media streaming. Mahanti et al. [28] did a comprehensive study on the file hosting ecosystem over a one-year period and concluded that the economic model based on advertisement and subscription revenue is sustainable. The optimal advertising level for a site is found to be closely associated with the advertising effectiveness, revenue discount rate, and the advertising-to-sales ratio [22]. Dewan et al. [9] consider the disutility from advertising and show that it may be optimal for a Web site to initially deploy fewer advertisements and more content. Fan et al. [11] develop a model for optimal revenue allocation when offering both subscription and advertising options to users and find that the advertising level should be kept low as advertisement rate increases. Tåg [38] finds that a firm offering an ads-free option would increase the advertising level for its ads-supported option, in turn making the ads-free offering more benefit to the content provider and advertisers but not consumers.

2.2. Versioning

Service versioning is a marketing strategy that provides levels of a service at different fare points. For example, Hosanagar et al. [18] examine best-effort service and premium service to determine the optimal pricing and capacity allocation policies for an Internet access provider. Their results show the viability of offering the best-effort service free while charging the premium service. Mandjes [29] investigates priority queuing as a way to establish differentiated data service according to users' delay-sensitivity; however, Fishburn and Odlyzko [15] show that two-tiered service differentiation could satisfy higher demand but create lower revenue than a single-price network with high QoS for all users.

Literature on software versioning focuses on the *functional* aspect of software systems by studying vertical differentiation along pricing and feature sets [6], timing of software upgrade/patching [8], and developer's commitment to competing hardware platforms [30] when deciding system features. In addition, previous studies on HCI have addressed *non-functional* attributes such as usability and user experience in general [36]. Our main concern is an important non-functional attribute, namely, the system's performance and responsiveness.

2.3. Resource allocation

Liu et al. [27] trade off the benefit of delivering more personalized content with longer waits due to the customization overhead. They treat service time for each request as a decision variable and present an efficient scheduling policy based on batching. Huang and Sundararajan [19] investigate the optimal capacity planning when the cost of service capacity is discontinuous and declining over time to show that the widely-adopted full cost recovery policies are often suboptimal. Analyzing the financial impact of service quality, Gallaugher et al. [16] find that the revenues from advertising sales and subscription fees are positively associated with the responsiveness of the service site; the provider thus should process requests quickly. Liu et al. [26] consider an e-tailor's site promotion problem when encountering IT capacity constraint in a duopolistic setting. Despite the increased traffic from higher advertising expenditure, such spending could render ineffective if congestion delay leads to customer attrition.

Several studies [10,12,37] improve the allocation of service resources by incorporating queuing modeling; some recent studies [1, 17,41] also treated the issue of satisfying user demands for given spending limits on cloud computing resources. While some studies [10,12,37] share the feature of queuing analysis, literature related to the capacity investment decision under data rate and advertising programs remains sparse despite the rapid expansion of ad-supported cloud services. Even fewer studies to date explicitly trade off advertising revenue and subscription fees under the lens of queuing policies.

Our approach investigates how service prioritization affects the providers for online storage and media streaming services when their revenue model includes both subscription and advertising. The revenue implications from service priority and differentiated congestion delay developed by Cheng et al. [5] (in terms of FCFS and priority scheduling) and advertising revenue developed by Prasad et al. [33] (in terms of adverting levels and subscription fees) are major influences for the current work. In turn, by incorporating both revenue models and service policies, our contribution to the literature is to derive the optimal capacity allocation and advertising levels under different ads rendering and service scheduling policies.

3. The model

Consider a provider that offers tiered service quality and features. For example, a cloud storage service provider[9] offers its premium

[6] http://www.fiercebroadbandwireless.com/story/report-mobile-data-traffic-patterns-look-similar-fixed-broadband-patterns/2010-03-21, (2010).

[7] YouTube to launch music streaming service, take on Spotify http://tech.fortune.cnn.com/2013/03/05/youtube-streaming/.

[8] Online Data Helping Campaigns Customize Ads http://www.nytimes.com/2012/02/21/us/politics/campaigns-use-microtargeting-to-attract-supporters.html?pagewanted=all&_r=0.

[9] For example, http://www.filefactory.com/ and https://www.rapidshare.com/.