Web Access Consolidation and Access Pattern-Based Optimization for Web-based Service Platforms

At the Research Laboratory we are conducting research into the use of Web-based Service Platforms in cloud infrastructure as part of research and development for next-generation cloud solutions. In this report we discuss Web server optimizations based on the consolidation of Web access and access patterns, which are the key to achieving high scalability and resource utilization for Web-based Service Platforms.

3.1 Web Server Resource Design

Typical Web server construction involves forecasting website request numbers and traffic volume and estimating the necessary content throughput, before estimating system resources and designing the server configuration based on this data. Web access variation is taken into account when making these forecasts.

There are two main variation patterns for Web access. The first is a cyclical variation pattern that occurs on a daily or weekly basis. The second is a variation pattern based on concentrated access. Concentrated access can be caused by a variety of factors. Some examples include news releases, game or software releases, links from the top page of major news sites, and URL captions on TV.

Because Web access can surge suddenly due to concentrated access, variations in Web access are taken into consideration during Web server resource design and backup resources added to estimates (over-provisioning). As shown in Figure 1, when implementing resource design using over-provisioning, the more backup resources that are added to the estimate in preparation for concentrated access, the lower the utilization of Web server resources will be.

Because Web services account for a large percentage of data center and cloud services, improving the resource utilization of Web servers also leads to improved overall resource utilization for data centers and cloud solutions.

3.2 Improving Resource Utilization through Web Access Consolidation

Many websites experience concentrated access following events with significant social impact such as the recent Great East Japan Earthquake. However, it is generally accepted that concentrated access will occur at different times for individual websites. This is because the causes of concentrated access and extent of its impact are both limited. Assuming that concentrated access occurs randomly at individual websites in accordance with probabilistic distribution, by consolidating traffic for multiple websites it stands to reason that the probability of concentrated access occurring over this consolidated web access would increase in proportion to the number of consolidated services. We therefore envisaged that it would be possible to reduce the overall variation in Web access by exploiting the fact that consolidated Web access raises the probability of concentrated access occurring. In other words, this would mask the variance in Web access due to concentrated access by intentionally creating a Web access environment that constantly faces concentrated access.

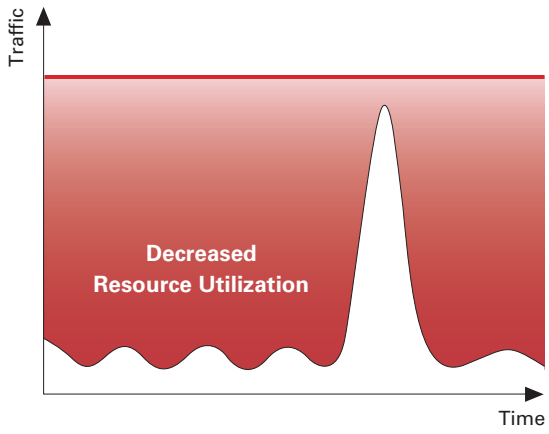

Figure 1: Decreased Resource Utilization through Over-Provisioning

To verify this concept we simulated a consolidated Web access environment in which concentrated access occurred at random, and observed the impact of individual pockets of concentrated access as well as the overall variance in access volume. The results are shown in Figure 2. The bottom graph shows the variance in Web access volume when Web services are not consolidated. The other three graphs show the variance in Web access volume when 10, 100, and 1,000 websites are consolidated, respectively.

The sharp spikes in these graphs indicate concentrated access. To make it easier to see the variance in access volume due to concentrated access as well as its impact, we have not included variance patterns for cyclical Web access.

The simulation results showed that consolidating Web access during concentrated access that occurs on a sporadic basis at individual sites led to comparatively less overall impact from access variance due to individual pockets of concentrated access. From this we believe it will be possible to smooth out Web access variance through the consolidation of Web access.

Next we will examine the merits of consolidating Web access with regard to resource design. Because consolidation masks and smoothes out the access variance resulting from concentrated access, resource design for consolidated websites can be implemented based on Web access with less variance. As shown in Figure 3, we believe this means that it will be possible to free up backup resources previously retained on an individual website basis, and lower the overall estimated backup resources required. For this reason we believe that it will be possible to improve the overall resource utilization of websites by consolidating them and basing the resource design on this.

3.3 Access Variance for Popular Content

We believe that higher efficiency for Web servers will be made possible through improving the processing efficiency of requests to consolidated Web access environments. We are looking at adapting request processing to Web access patterns as one method of implementing this.

Because concentrated access occurs on the Web due to certain limited factors, the content where access is concentrated is also limited. For this reason, when Web access is consolidated access will be concentrated on a fixed percentage of specific content. It should be possible to improve Web server optimization through the efficient processing of requests for the content where access is concentrated.

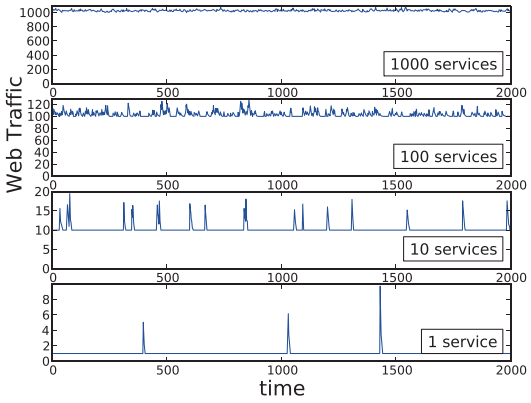

Figure 2: Differences in Web Access Variations due to Website Consolidation

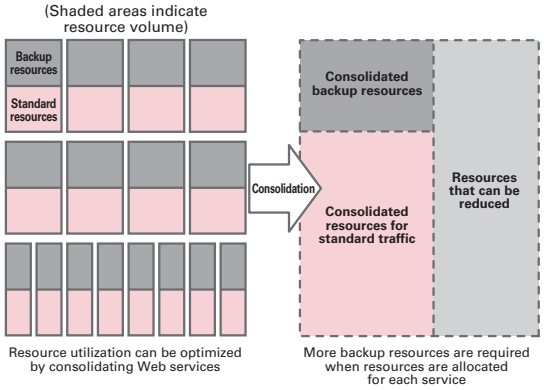

Figure 3: Backup Resource Reduction due to Website Consolidation

The content where access is concentrated is not always the same, and will shift over time. To process these requests efficiently it is necessary to consider the degree that access is concentrated on a given of piece content as well as changes in this concentration. To this end we analyzed content where access was concentrated based on the actual traffic logs of Web servers. For this analysis we used traffic logs for the Web servers of the Japan Advanced Institute of Science and Technology (JAIST). JAIST operates the largest Web service for the distribution of free software in Japan, resulting in a large number of users converging there. It is also the largest domestic distribution site for Firefox, so concentrated access can be predicted to accompany software releases.

Figure 4 is a graph indicating the cumulative distribution function (CDF) of content, traffic volume, and requests in order of peak requests for each piece of content based on a random sampling of 17,000 pieces of content, accounting for approximately 2% of the content requested over the course of a certain month at JAIST. Traffic and requests are the total number for each piece of content over the period of analysis (180 days). From the cumulative distribution function of content (the green dashed line) in this figure, we can see that approximately 90% of the peak requests for content amounted to less than 10^2 (100) requests per day. However, looking at the cumulative distribution function for traffic (red line) and requests (blue line) in the figure, it is clear that content with peak requests of less than 10^2 (100) requests per day accounted for a tiny percentage of approximately 2% of traffic and requests. This points to the fact that access is skewed towards content with more than 10^2 (100) peak requests per day.

Based on these results, we analyzed changes in the concentration of access for content receiving more than 200 requests per day. We identified the number of requests per day for each piece of content over the analysis period of 180 days, and labeled the highest value as the day that requests peaked. Additionally, in order to analyze the variation in requests for a period of 10 weeks after peak, we elected to use only data for which the day that requests peaked was at least 70 days before the end of the analysis period. As a result, we gathered data on approximately 3,000 pieces of content meeting the criteria.

Figure 5 and Figure 6 show the peak requests as well as changes in the number of requests for each piece of content 1 day and 7 days after requests peaked. The horizontal axes of these figures show the number of requests for the content on the day that requests peaked. The closer to the green line, the less change was observed in the number of requests after requests peaked. Content for which the number of requests fell to 0 either 1 day or 7 days later were recorded as 10^{-1}.

The spread of red dots on these figures indicates the drop in the number of requests compared to the peak requests. Comparing results for 1 day and 7 days later, the spread of points is wider for 7 days later, showing that for most content the more days that pass after requests peak, the greater the decrease in requests will be. Following the same

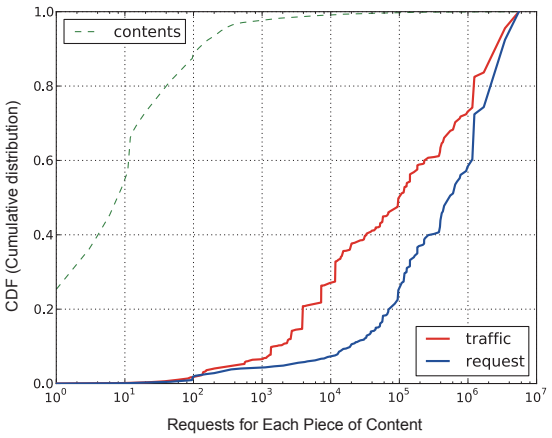

Figure 4: Content, Requests, and Traffic Share based on Peak Requests for Each Piece of Content

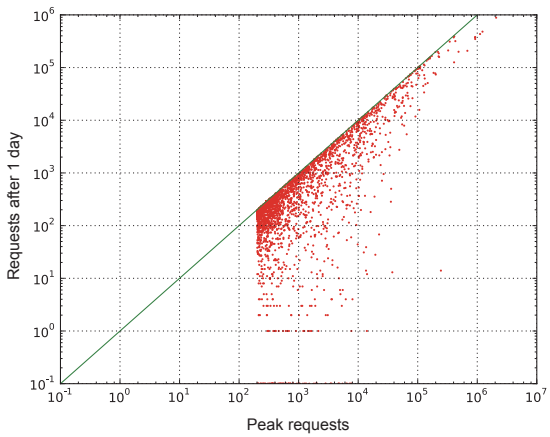

Figure 5: Changes in Requests for Each Piece of Content (1 Day Later)

method, we compared changes in the number of requests over 10 weeks on a weekly basis. In this case the spread of points stopped gradually, showing that after 4 weeks requests continued to be received for a certain amount of content even when 10 weeks passed.

Using the same data, we also identified the rate of decline in requests from the day that requests peaked to 10 weeks later for each piece of content. Figure 7 shows the mean value (blue line) and median value (red dashed line) for the rate of decline in requests. From this figure, we can see that the number of requests was approximately 40% of the peak value about 1 week after the day that requests peaked, and this value continued to decline gradually thereafter. The slope of the median value is greater than that of the mean value, and approaches 0 first. In other words, it appears that the number of requests is declining faster than the average value for most content because the decline in requests is more gradual for some content, pushing the mean value up.

It is likely that the numerical results obtained here are unique to JAIST Web servers and do not apply to general Web services. However, we can consider this one pattern of variation in requests for content where access is concentrated. We believe that when consolidating Web access where concentrated access occurs on a constant basis, analyzing the behavior surrounding content such as this is crucial for evaluating the optimization of request processing on Web servers.

3.4 Conclusion

The information on Web access consolidation and optimization of request processing adapted to content access patterns presented here is currently still at the research stage, and we do not yet provide services based on these design concepts. However, this does not change the fact that Web services will continue to play a key role in information infrastructure. We believe that research and technological development aimed at expanding the scope of Web services using large-scale resources such as cloud computing will become more important in the future. We plan to continue this research as part of larger research goals such as these.

In closing, we would like to offer our sincerest thoughts and prayers to those who lost their lives in the recent earthquake. We pray for the swift recovery of those affected by this tragedy.

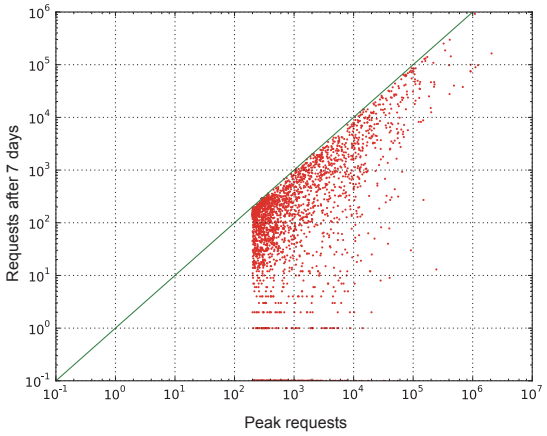

Figure -6: Changes in Requests for Each Piece of Content (7 Days Later)

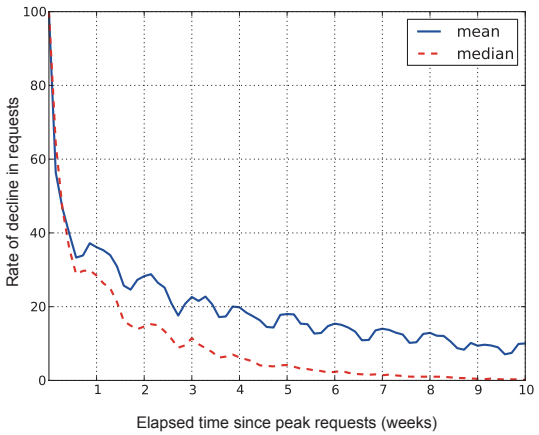

Figure 7: Rate of Decline in Access Numbers for Each Piece of Content

Author:
Megumi Ninomiya
Research Laboratory Research Associate, IIJ Innovation Institute Inc. Ms. Ninomiya is engaged in the research of Web-based Service Platforms with the goal of achieving high performance and high resource utilization for cloud infrastructure.