Understanding Service Demand for Adaptive Allocation of Distributed Resources

Jose Renato Santos, Koustuv Dasgupta[1+], G. (John) Janakiraman,
Yoshio Turner
Internet Systems and Storage Laboratory
HP Laboratories Palo Alto
HPL-2002-85
April 8th , 2002*

Internet services experience frequent changes in demand, resource characteristics and service requirements. A service can meet its requirements with minimal cost only if the allocation of its distributed resources can be adaptively controlled. The design of adaptive control schemes requires a solid understanding of the dynamic characteristics of the distributed service, demand and resources. This study examines the dynamic properties of distributed demand and differs from prior demand characterization work by focusing on dynamic variations across clients, time and region which are crucial in adaptively allocating distributed resources. Our analysis of the demand for the 1998 World Cup web site finds that the dynamic behavior of a small subset of clients is representative of the entire demand, the churn in the active set of clients from day to day is relatively small and that regional demand shows significant, and predictable, variations in the hourly scales. These results will provide guidance for the design of adaptive policies.

Understanding Service Demand for Adaptive Allocation of Distributed Resources

Jose Renato Santos[‡], Koustuv Dasgupta[§¶], G. (John) Janakiraman[‡], Yoshio Turner[‡]

[‡] Hewlett Packard Laboratories
1501 Page Mill Road
Palo Alto, CA 94304

[§] University of Maryland at Baltimore County
Dept. of Computer Science and Electrical Eng.
Baltimore, MD 21250

Abstract— **Internet services experience frequent changes in demand, resource characteristics and service requirements. A service can meet its requirements with minimal cost only if the allocation of its distributed resources can be adaptively controlled. The design of adaptive control schemes requires a solid understanding of the dynamic characteristics of the distributed service, demand and resources. This study examines the dynamic properties of distributed demand and differs from prior demand characterization work by focusing on dynamic variations across clients, time and region which are crucial in adaptively allocating distributed resources. Our analysis of the demand for the 1998 World Cup web site finds that the dynamic behavior of a small subset of clients is representative of the entire demand, the churn in the active set of clients from day to day is relatively small and that regional demand shows significant, and predictable, variations in the hourly scales. These results will provide guidance for the design of adaptive policies.**

I. INTRODUCTION

The infrastructure supporting many Internet services and the clients of these services are globally distributed (e.g., Olympics96 [1], WorldCup98 [2], CDNs [3] [4]). Since the client population is practically unbounded for many services, demand for these services can vary over a wide range with peak demand one or two orders of magnitude larger than average demand. It is not cost effective to statically provision the service infrastructure with enough resources to meet service level requirements under the worst-case demand. Therefore, services that must meet specific service level objectives must be deployed on an adaptive infrastructure whose resources can be scaled dynamically in proportion to instantaneous demand.

An adaptive infrastructure for distributed services must have the capability to dynamically control several levers including 1) the distribution (or placement) of service sites, 2) the partitioning of service functionality among the service sites, 3) the amount of resources allocated for each site and 4) the assignment of client requests to these sites. Many of the underlying mechanisms required to realize an adaptive distributed service infrastructure are available (at least in rudimentary form) today. Examples include means to snapshot a service instance for migration or replication to another site (e.g., Ejasent's Instant Application Switching [5]) and means to automatically adjust the number of resources allocated to the service at a data center (e.g., HP's Utility Data Center [6]).

To facilitate the design of an effective adaptive distributed service infrastructure, it is essential to develop a good understanding of the dynamic variations in the service demand. While several previous studies [1], [7], [2], [8], [9] characterize demand for Internet services, they are primarily useful in designing a single site for the service and in caching studies. They do not examine workload properties that are crucial in adaptively allocating distributed resources. Other studies [10], [11] cluster clients based on their topological proximity to evaluate placement algorithms. However, they do not examine the regional distribution of these clusters.

In this paper, we analyze temporal variations of demand across the network topology, using the 1998 World Cup web site workload [2]. Understanding the topological distribution and temporal variation of demand is important for allocating distributed resources such that latency between clients and the service, as well as the cost of service infrastructure is minimized. The intensity of demand for this web site has been shown to vary substantially over the days of the World Cup event [2]. We extend that previous work by focusing on the distribution and temporal variation of demand across the clients. In addition, we analyze the regional distribution of clients and how demand is distributed across network regions over time, studies that to our knowledge are unique. The findings provide insight needed to design policies for adaptive distributed resource allocation.

II. RELATED WORK

Solutions to adaptively allocate resources at a single site are being developed by many system vendors. HP's Utility Data Center [6], IBM's Oceano [12], and Peakstone [13] all support means to dynamically allocate resources in a single site among multiple customers sharing the site. CDNs such as Akamai [3] and DigitalIsland [4] host services on a geographically distributed infrastructure. Ejasent [5] also hosts services on a distributed infrastructure but in addition provides tools for dynamically deactivating and reactivating an application on different sites. However, details of resource allocation policies and mechanisms both for Ejasent and CDNs are proprietary and not publicly available.

Previous studies have characterized demand for various Internet services including [1], [7], [2], [8], [9]. Most of these ([1], [7], [2]) primarily analyze the aggregate demand which is

---

[¶]Work done during internship at Hewlett Packard Laboratories

useful in examining Web server performance for capacity planning of a single site. They do not address issues that are important for designing an infrastructure with distributed resources. Other studies [8], [9] examine content usage characteristics in the demand for the purpose of evaluating caching and prefetching mechanisms and algorithms.

Krishnamurthy and Wang [10] proposed a technique for clustering clients (which we describe and use in Section IV) and used it to evaluate the benefit of placing proxy caches on these clusters. Recently, Qiu et al described algorithms for placing web server replicas in the network in [11]. They derive load information for their evaluation by applying the same client clustering technique to an MSNBC server workload. One of their conclusions is that their placement algorithms perform well even when they are based on cluster information from the past. We derive a similar conclusion in Section IV for a different workload. However, Qiu et al [11] do not consider the network topology associated with their experimental workload, while we investigate the actual regional distribution of the load and the temporal variations in this regional demand.

III. The Dataset

Our study of distributed demand uses the log of all accesses to the France 1998 World Cup web site [2]. We chose this data set for two reasons. First, among the datasets with client information available to us, this had the heaviest load. Second, the World Cup event had world wide interest creating a global workload with clients distributed all around the world.

The World Cup web site provided information such as real-time match results, historic data on previous games, player statistics, team history and news articles. The tournament lasted for 33 days. The dataset consists of the log of all web accesses to the web site during 88 days, starting 41 days before the tournament began until 14 days after the tournament ended In this period, the web site received 1,352,804,107 requests for 4,991 GBytes of data from 2,770,108 unique clients (IP addresses). More details about the workload can be found in [2].

IV. Workload Analysis

In this section we analyze characteristics of the 1998 World Cup workload that are important for global resource allocation. We evaluate how the demand is distributed across the client population, how the client population changes with time and how the demand is distributed over large topological regions of the worldwide Internet.

The large number of clients in the dataset makes it difficult to analyze the demand characteristics based on the raw client data. For the same reason, interpreting the results of such an analysis would also be hard. It is desirable to group the clients into a smaller set of client groups to facilitate analysis and interpretation. However, it is critical to preserve the actual topological distribution of clients in this grouping since this factor is key in distributed resource allocation.

A solution to this problem is to cluster clients in non-overlapping groups based on their topological proximity. Krishnamurthy and Wang [10] proposed such a a clustering technique which uses information available on BGP (Border Gateway Protocol) routing tables, to cluster IP addresses. We adopted their technique, which we refer to as BGP clustering in the rest of this paper. By examining multiple BGP routing tables the BGP clustering technique can identify IP address ranges that share the same routes from multiple places and thus are likely to be topologically close to each other. In [10], the authors validates the BGP clustering technique using both a "traceroute" approach and a "domain name" approach, showing that more than 90% of the clusters they had identified have 100% of their clients topologically close to each other.

We applied the BGP clustering technique to all clients that visited the World Cup web site, using BGP table snapshots obtained from [14], [15], [16], [17], and [18]. Based on these BGP tables we clustered the 2,770,108 unique client IP addresses found in the web site log into 81,420 unique clusters. The number of clients in each cluster varied from 1 to 16,539, with more than 95% of clusters having fewer than 100 clients and more than 87% of the clusters having fewer than 20 clients.

Fig. 1 shows the daily demand to the World Cup web site, showing the number of HTTP requests and the number of *active* clusters (i.e. the number of unique clusters that accessed the site) on each day of the 88-day period. We observe that an increase in load is associated with an increase in the number of active clusters accessing the site, as expected. We observe that traffic demand presents huge variations (confirming previous results [2]), which suggests that dynamic allocation of resources is beneficial. Note that the period of higher demand coincides with the World Cup event which started on day 42 and ended on day 74.

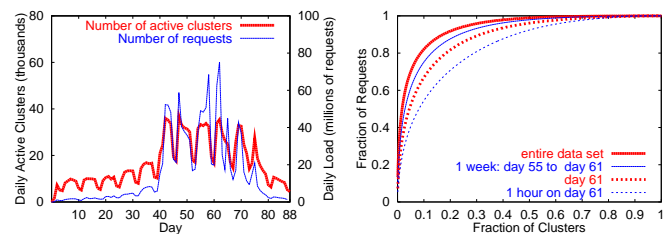### A. Demand distribution across clusters

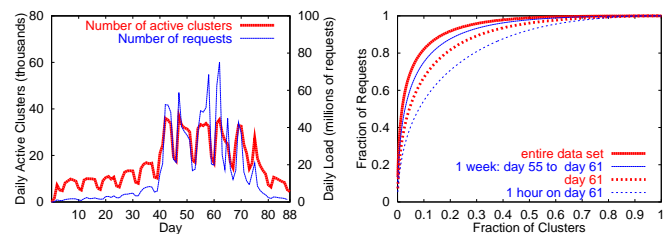Fig. 1.   Daily demand   Fig. 2.   Skewed Demand

Fig. 2 shows the distribution of demand across the clusters. The graph shows the normalized cumulative load as a function of the fraction of the clusters, selected in decreasing order of load. Each curve in the graph corresponds to load generated over a different time period (the entire dataset, a week, a day, and an hour). For each curve, the load is normalized by the total load during the period and the fraction of clusters is computed considering only clusters that were active during that period. The graph shows that the load is highly skewed, with a small fraction of the clusters being responsible for a high fraction of the load, confirming previous results [10]. We refer to these highly active clusters as *dominant* clusters. We also observe that the load is less skewed for shorter time intervals. This can be explained by the fact that clusters generating a small number of requests (referred to as *non-dominant* clusters) are less likely to
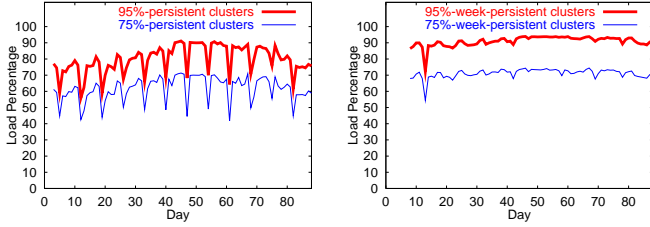
(a) p-persistent clusters   (b) p-week-persistent clusters

Fig. 3. Persistence of active clusters

(a) absolute load   (b) fraction of total daily load

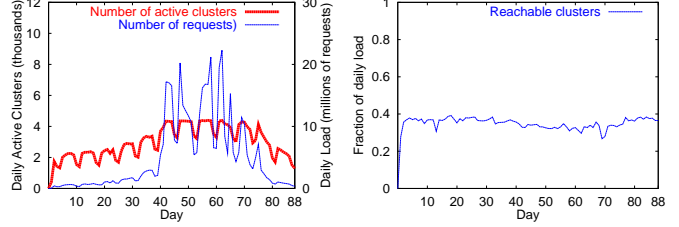Fig. 4. Daily demand for reachable clusters

be represented in smaller time periods. As a result, for the same fraction of the load, in smaller time periods, dominant clusters are a larger fraction of the active clusters, which corresponds to a less skewed load. This skew can be exploited to reduce the cost of making resource allocation decisions. These decisions can be based on measurements for the small subset of dominant client clusters since they generate most of the workload.

B. Predictability of dominant clusters

In order to examine the predictability of the cluster population accessing the service, we define *p-load* and *p-persistent* clusters. We define *p-load* clusters for any day as the most dominant clusters of the day which are responsible for *p%* of the total load on that day. We define *p-persistent* clusters for any day as the ones which are part of the *p-load* clusters on that day and were also part of the the *p-load* clusters on the previous day, i.e. the intersection of the *p-load* clusters of two consecutive days. Fig. 3(a) shows the fraction of the daily load generated by the *95%-persistent* and *75%-persistent* clusters during the entire dataset period. For the *95%-persistent* case, we observe that except for periodic low spikes, 80% to 90% of the total load is associated with clusters that were also dominants in the previous day. Similarly, for the *75%-persistent* case, 60% to 70% of load was generated by persistent clusters. This result indicates that the dominant client population is highly predictable. Thus, for this dataset, it is possible to achieve good resource allocation decisions, particularly placement decisions, based on historical data of client population. Qiu et al [11] observed similar behavior for their dataset.

The low spikes in the *p-persistent* curves of Fig. 3(a) occur every Monday and are due to the load difference between weekdays and weekends. The demand for the World Cup web site was significantly lower during weekends than during weekdays. As reported in [2], this is assumed to be a consequence of the fact that more people were able to follow the events through TV coverage during weekends. Due to a higher daily load, the *p-load* set has more clusters on Monday than on Sunday, generating an intersection set (i.e. the *p-persistent* set) with a smaller fraction of the Monday clusters which in turn corresponds to a smaller fraction of the load. This result indicates that high fluctuations in the load may reduce the predictability of the client population from the previous day.

One way to improve predictability of future client population is to use several days of history information. Fig. 3(b) shows the load associated with persistent clusters using 7 days of previous history. We define the *p-week-persistent* clusters as the

p-load clusters in a particular day which were also part of the of *p-load* clusters in at least one of the 7 previous days. As we observe in Fig. 3(b) the spikes are significantly reduced when *week persistence* is used. We also notice that the load associated with the persistent set increases. We see that the "new" clusters which could not be predicted from the previous 7 days were responsible for approximately only 5% of the load. As mentioned before, this high predictability of the client population can be exploited for accurate resource allocation decisions.

C. Regional demand

We now explore regional demand characteristics of the World Cup web site. Our goal is to extract characteristics inherent in the distribution of demand across the worldwide Internet such as the network distances of clients from various global sites, regional clustering of clients and temporal variations across regions. These characteristics provide information on the network latency to potential service sites and the potential load on each service site. Global resource allocation decisions such as the "best" placement of service sites and the "best" assignment of client requests to service sites can be based on this information. Other factors such as cost of site installation, flexibility to dynamically allocate resources and maximum capacity on each site must also be considered when designing adaptive resource allocation policies.

To characterize the network distances of clients from various global sites, we use several public "ping servers" available on the Internet and measure client latencies to these ping sites. A "ping server" is a web site that allows users to measure network latencies from the server site and an arbitrary IP address. The ping server sends ICMP echo messages to the specified IP address and reports the measured average round trip time through a standard HTML page. In all experiments used in this paper the round trip time was computed as the average over 10 measurements. We used 17 ping servers located as follows: 6 in the United States, 6 in Western Europe, 2 in Eastern Europe, 1 in Canada, 1 in Australia, and 1 in Africa.

As we mentioned earlier, it is impractical to analyze and interpret measurements of each individual client with such a large client population. Since the BGP client clustering technique preserves the topological proximity of clients, the round trip times to the clients of any one cluster from an arbitrary point in the Internet will all approximately be the same. Thus, the round trip measurement for any one client in a cluster can be used for all clients in the cluster.
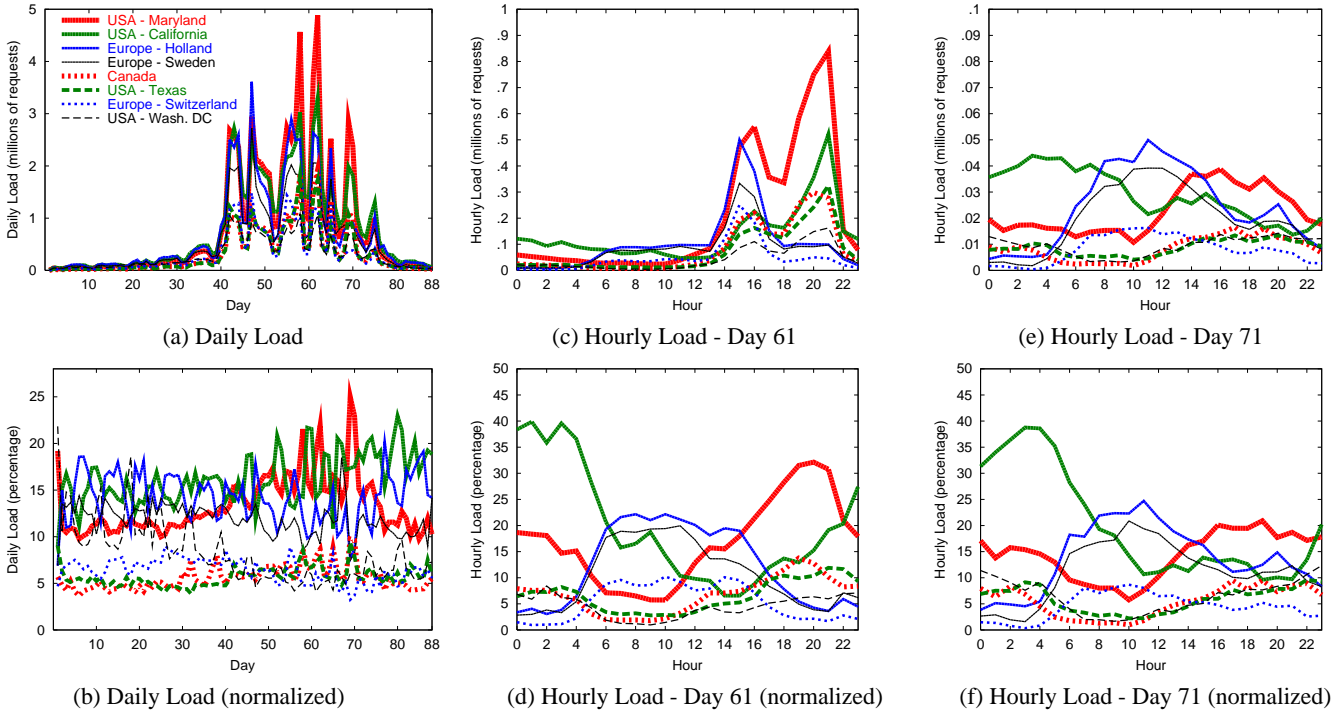
Fig. 5.   Regional Demand

Due to the large number of clusters (81420), for the results presented in this section, we considered only the most active clusters responsible for 90% of the load during the entire period. This reduced the number of clusters used in round trip time measurements to 14053. For each of the 17 ping servers, we measured the average round trip time from the server location to each of these most active clusters. This was done by choosing a random client belonging to each selected cluster and submitting its IP address to the associated ping server. Unfortunately many clusters were not reachable at the time of our experiment. We classified a cluster as non reachable after 5 unsuccessful measurements to 5 different random clients belonging to that cluster. Of the 14053 selected clusters, only 4456 were reachable in our experiments. The load associated with the reachable clusters is shown in Fig. 4. We observe that the daily load pattern for the reachable clusters is almost indistinguishable from the total load shown in Fig. 1, except that the absolute magnitude was reduced to approximately 40% of the original value, consistently over all days. This gives us confidence that the reachable clusters constitute a representative sample of the clusters and our results could be extended to the overall population.

As expected, our experiments showed a wide variation in the round trip latencies across the clients (not presented here due to space limitation). Latencies vary from less than 10 milliseconds to more than 10 seconds depending on the topological proximity of the client and the server. These large variations in network latency indicate that judicious placement of resources is important, at least for applications that require short response times.

Since the ping servers are globally distributed, we use the measured round trip times to group the reachable clusters into 17 groups, corresponding to each of the 17 ping servers. Each cluster was assigned to the "closest" ping server based on its round trip measurements. This assignment corresponds to having all 17 candidate sites hosting the service and routing clients to the closest site. We refer to each of these 17 groups of clients as one region.

Fig. 5 shows the demand associated with the 8 most active regions. These regions were responsible for approximately 80% of the total load generated by the reachable clusters during the entire period. The other 9 regions had small loads varying from 0.15% to 5% and are, hence, not shown in Fig. 5 for clarity. Fig. 5(a) shows that, in general, each region presents the same pattern as the aggregate load shown in Fig. 1. This is confirmed by Fig. 5(b) which shows the load of each region, normalized by the total load on each day. Although there is some oscillatory behavior, on average, each region is responsible for approximately the same fraction of the load on all the days. This is true even when we compare days with large differences in absolute load, as for example the days during the World Cup event compared to the days preceding the event. This fact suggests that it is possible to predict load distribution across multiple topological regions from historic data, even if the total absolute load cannot be predicted. Therefore, day to day changes in the location of resources may not be required (for this dataset) purely to deal with regional shifts in load. However, note that other factors such as regional data center cost variations or changes in total demand may still necessitate dynamic placement of resources. For example, if a service experiences a surge in total demand such that one of the regional servers experiences an increase in demand beyond its maximum capacity, service placement decisions would have to be dynamically re-evaluated.

Although the daily distribution of load is relatively constant,

there is more variation when the load is examined at a finer time resolution. Figs. 5(c) and 5(e) show the load on a time scale of hours for the same regions shown in Figs. 5(a) and 5(b), on days 61 and 71 respectively. Figs. 5(d) and 5(f) show the normalized loads for these two days, respectively. We selected these two days because the load behavior differs markedly on these days. There were two games played on day 61 which caused a relatively high load on that day. Also, the load was significantly higher at game times than during other times of the day. The two peaks on the curves in Fig. 5(c) coincide with the times of the games as reported in [2]. On the other hand, there were no games played on day 71, and as expected, the load on this day was much lower (Note that the scale of the y-axis for Figs. 5(c) and 5(e) are different). The load was also relatively more uniform throughout the day. From Figs. 5(d) and 5(f), we observe that the hourly distribution of load changes significantly during the day. The set of clusters dominant during any hour changes several times during the day (each regional site corresponds to a different subset of clusters).

However, comparing Figs. 5(d) and 5(f), we observe that there is a similar pattern for the normalized load for these two days, despite the completely different absolute load patterns. Each region is responsible for approximately the same fraction of the load at the same time of the day, on both days, but this fraction changes with the time of the day. This behavior can be explained by the time zone differences among the different regions. For example, the curves for Maryland and California in Fig. 5(d) present peaks in load approximately 5 hours apart, which is close to the 3 hour time difference between these two locations. The distance between the load peaks do not match the time zone differences exactly probably because each candidate site may be assigned clients from many different time zones. For example, clients from Central and South America may be assigned to United States sites.

These results suggest that time zone differences do have an impact on the load distribution. If the service characteristic and data center infrastructure would allow fast and cost effective deployment and shut down of service sites, dynamic service placement could be beneficial at an hourly time scale. Moreover, since the hourly load distribution pattern is predictable from the previous day, dynamic service placement could be done in a planned manner.

Note that load distribution based on network proximity alone can lead to non-uniform distribution of load across sites. Certain sites may see much higher demand than other sites. This distribution can also change with time (e.g., a site may be busy during the day and idle at night). In an environment with dynamically shared resources, these differences may translate into differences in cost and supply of resources. Adaptive resource allocation policies will have to balance these factors against network proximity requirements. For example, batch applications may tolerate longer latencies, and may take advantage of lower cost resources at locations with current low demand.

V. CONCLUSION

We studied the characteristics of the demand to the 1998 World Cup web site to provide us guidance in developing adaptive methods for distributed resource allocation. We find that the small subset of client clusters (clustered using the technique in [10]) which dominate the demand is stable on a day to day basis. On an hourly timescale within a single day, however, the active set of clusters and their proportion of the hourly demand changes significantly, albeit predictably from day to day. Hence, dynamic adjustments to the allocation of distributed resources several times during the day would be worthwhile. This could be accomplished by changing placements and/or by scaling (up and down) the number of resources at each site depending on factors such as the difference between the costs of placement changes and resource scaling and differences in resource capacity and cost across sites.

Many Internet services may exhibit very different demand behavior than the 1998 World Cup site. For example, demand for streaming media services, services for mobile clients or online multi-player game services may have very different characteristics due to differences in their content attributes (e.g., popularity, reuse and update frequency), client attributes (e.g., physical relocation, network capabilities) as well as the nature of the service (e.g., dynamic processing). We intend to investigate the implications of the demand for such services on adaptive resource allocation. As we mentioned earlier, adaptive policies will also be dependent on variations in resource behavior such as changes in network latency and bandwidth. With this additional insight, we are interested in exploring the design of an adaptive infrastructure where placement, load distribution, resource scaling and function partitioning are controlled in an integrated way to exploit trade-offs across their solutions.

VI. ACKNOWLEDGMENTS

REFERENCES

[1] A. Iyengar, M. S. Squillante, and L. Zhang, "Analysis and characterization of large-scale web server access patterns and performance," *World Wide Web*, vol. 2, no. 1-2, pp. 85–100, 1999.

[2] M. Arlitt and T. Jin, "Workload characterization of the 1998 World Cup web site," Tech. Rep. HPL-1999-35, HP Labs, Sept. 1999.

[3] Akamai, "http://www.akamai.com".

[4] Digital Island, "http://www.digitalisland.net/services/cont_delivery.shtml".

[5] Ejasent, "http://www.ejasent.com/platform.shtml".

[6] HP, "http://www.hp.com/solutions1/infrastructure/solutions/utilitydata/overview/index.html".

[7] M. F. Arlitt and C. L. Williamson, "Internet web servers: Workload characterization and performance implications," *IEEE/ACM Tr. Networking*, vol. 5, no. 5, pp. 631–645, 1997.

[8] L. Breslau, P. Cao, L. Fan, G. Philips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *IEEE INFOCOM*, Mar. 1999, pp. 126–134.

[9] V. N. Padmanabhan and L. Qiu, "The content and access dynamics of a busy web server: Findings and implications," Tech. Rep. MSR-TR-2000-13, Microsoft Research, Feb. 2000.

[10] B. Krishnamurthy and J. Wang, "On network-aware clustering of web clients," *Computer Networks and ISDN Systems*, vol. 17, pp. 1–14, Aug. 2000.

[11] L. Qiu, V. Padmanabhan, and G. Voelker, "On the placement of web server replicas," in *IEEE INFOCOM*, Apr. 2001.

[12] K. Appleby, S. Fakhouri, L. Fong, G. Goldszmidt, and M. Kalantar, "Oceano - SLA based management of a computing utility," in *IFIP/IEEE Intl. Symp. on Integrated Network Management*, May 2001.

[13] Peakstone, "http://www.peakstone.com".

[14] "http://www.merit.edu/ ipma/routing_table".

[15] "telnet://route-views.oregon-ix.net".

[16] "ftp://rs.arin.net/netinfo".

[17] "http://moat.nlanr.net/ipaddrocc".

[18] "http://noc.singaren.net.sg/netstats/routes".