Variations in Cache Behavior.

Chris Roadknight and Ian Marshall

BT Research Laboratories, Martlesham Heath, Ipswich, Suffolk, UK. IP5 3RE

{roadknic,marshall}@drake.bt.co.uk

Keywords. Traffic Modelling, Fractals, Cache, Internet.

Abstract.

HTTP cache servers reduce network traffic by storing popular files nearer to the client and have been implemented worldwide. Their reported performance on key metrics such as hit rate varies greatly. In order to optimise the design of the cache network this variation needs to be understood. The variation in hit rate across a number of caches is investigated and is shown to be partly stochastic (i.e caused by insufficient sample size) and partly fractal (i.e deterministic in origin).

1.Introduction.

As the World Wide Web becomes more widely used, its performance as a large-scale distributed information system needs to be accurately modelled in order to optimise the system performance and minimise cost. Cache servers are a key component as they have been shown to improve performance and minimise network traffic, by caching frequently requested files close to the clients [ABR95] [BAE97]. However, cache behavior is difficult to model because cache performance indicators (i.e. hit rate, throughput rates) are highly variable. In order to construct accurate models the variation needs to be better understood.

2. Inter-cache differences.

This paper investigates inter-cache variation in hit rate and neglects differences in other parameters which vary in a similar manner. Effects due to differences in hardware and software were removed from enquiries by only considering caches with similar implementations and no hardware resource constraints such as disk space. In an attempt to maximise variation the caches investigated were chosen to represent a wide range of user communities. The caches investigated (all SQUID [WES96]) were:

NLANR (sv). A parent for many caches worldwide located at NASA-Ames/FIX-West in Silicon Valley

Edinburgh. Used by the staff and students of the University of Edinburgh, UK.

Human Genome Mapping Project (HGMP). Used by scientists working on the HGMP project in the U.K.

Korea. The Nowcom cache in Soeul, Korea.

ACT. Used by a number of educational institutions in Canberra, Australia.

JANET. The UK Academic and Research Community web cache.

Swinburne. Serves Swinburne university in Australia

Quest. Serves Queenslands education and science institutions.

Hit rates at caches exhibit a high degree of inter-cache variation [DUS97]. Figure 1 shows the mean and variance of the hit rate of 6 caches for October 1997. The caches investigated were found to have hit rates between 16% and 53% with a mean of around 30%. It is generally thought that cache hit rates seldom exceed hit rates of 50% even with infinite disk space [ABR95] [DUS97], and hit rates of less than 20% are also rarely seen, so the sample has captured a reasonable level of variation for further investigation

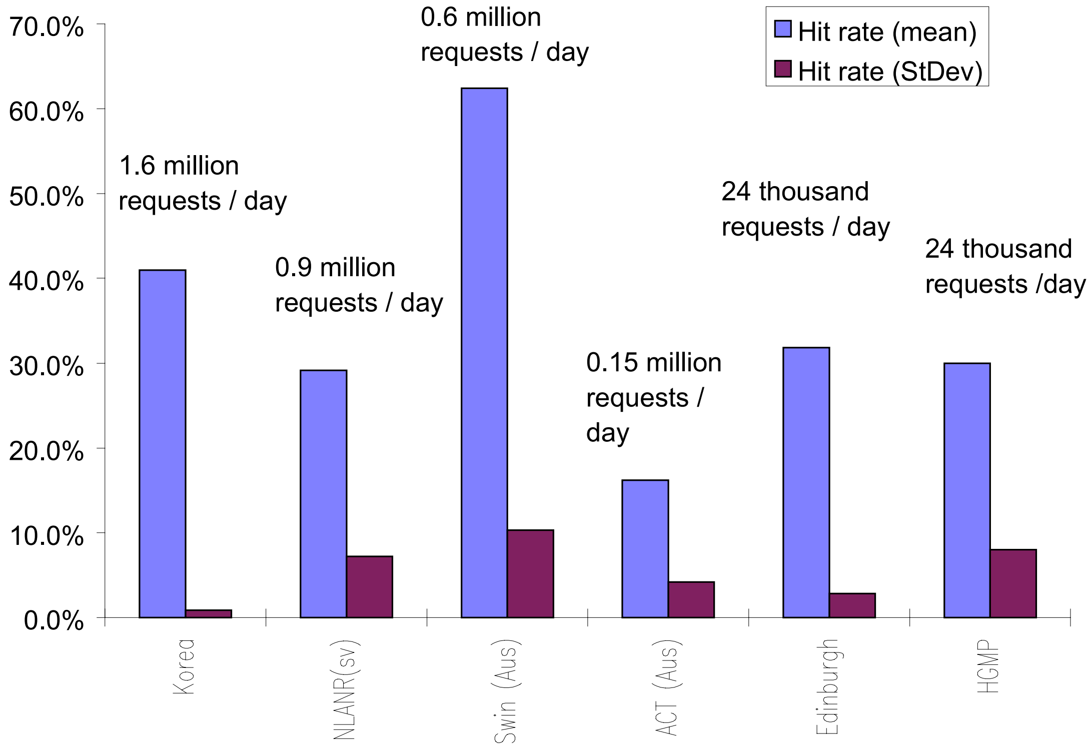

Figure 1. Hit Rate Statistics at 6 Proxy Servers.

The large difference in the variance of the hit rates for the 6 caches is remarkable, variance increases 60 fold between caches. Since the largest site has the smallest variance and the smallest site has the largest variance the high degree of variation in hit rates could be simply due to poor statistics, but the link between variance and cache usage does not appear to be strong.

The relationship between number of requests and variance in hit rate was further investigated by plotting variance divided by hit rate, against the number of requests received, for a range of sites (Fig 2). For all curves the leftmost point is averaged over one day and points to the right are averaged over successively longer periods up to 14 days.

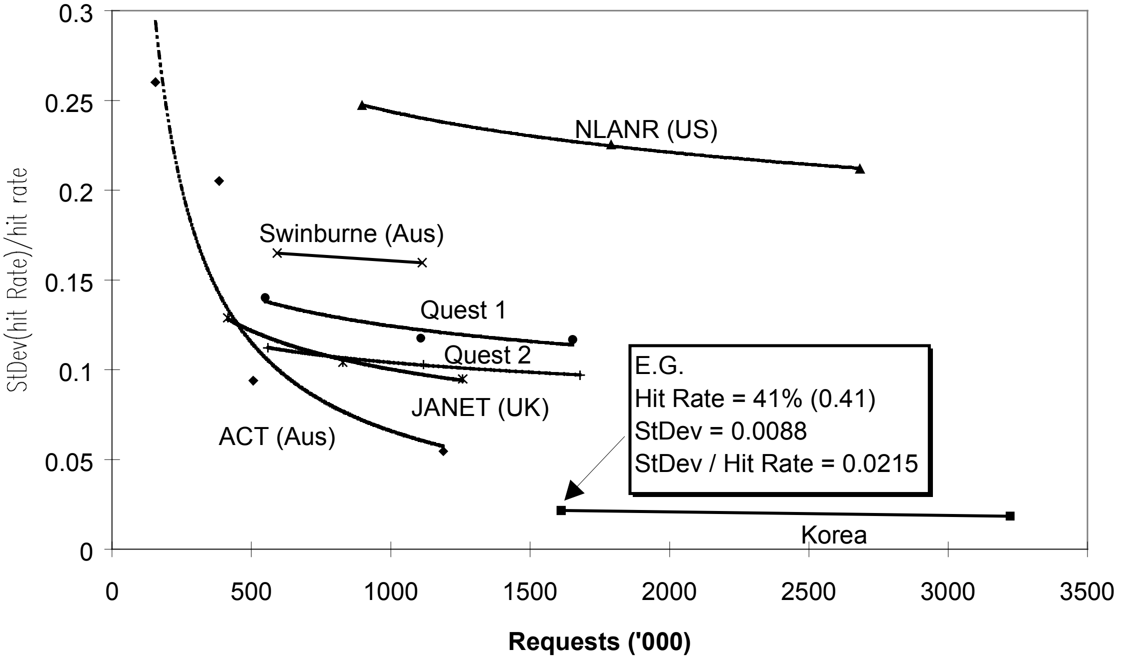

Figure 2. Effect of Number of Requests on Hit Rate Variation.

If cache performance variation were entirely stochastic, the variance would decrease exponentially with increasing sample size [WIL97]. It can be seen from figure 2 that this is only partly true - it appears that samples smaller than around 2 million requests are noisy. For larger samples the variance does not decrease strongly with increasing sample size. This indicates that the variation is self-similar over several time scales and is therefore fractal. This is a strong indicator that the underlying behaviour is chaotic. This should not be a surprise since the data is really an aggregation of the deterministic choices made by a large number of individuals. Other systems involving aggregation of this kind such as stock markets are already known to be chaotic [TAQ86].

In fact each cache exhibits unique behavior. The curves labelled Quest1 and Quest2 are derived from 2 servers that share the load from a single community. Since the load sharing is random, one would expect the two curves to converge as they do in the graph. However the Quest cache does not appear to be strongly convergent with NLANR or Korea. The systematic differences between caches are almost certainly the result of systematic differences between user communities, and indicate that any model will need to allow for a range of user types with different deterministic properties.

3.Discussion.

It has been shown in this work that a significant proportion of cache performance variation is fractal, the self-similar nature of the underlying network traffic [CRO96] supports this. The cause is probably the deterministic nature of users and user group behavior. Work to derive a model for these deterministic effects is currently in progress. It is expected that sophisticated modelling of user group behaviour will be required. In the future, client sets may be defined by their proportions of users with different usage requirements (Academic, Leisure, Work etc), culture (Oriental, Western etc.) and other variables. Their resulting request profiles may be predicted, based on these factors, using techniques such as artificial neural nets.

4.Conclusion.

Hit rate is important when analysing cache performance but there is a significant amount of intra and inter-cache variation. Some of the variation is stochastic in origin but much is fractal, i.e. deterministic in origin. Whether this is asymptotically self-similar, self-similar or multi-fractal is unclear. Further investigation is required to derive a model for the deterministic effects.

5.References.

[ABR95] M. Abrams, C.R. Standridge, G. Abdulla, S. Williams and E.A. Fox. Caching Proxies: Limitations and Potentials. Proc. 4th Inter. World-Wide Web Conference, Boston, MA, Dec. 1995. http://ei.cs.vt.edu/~succeed/WWW4/WWW4.html

[BAE97] M. Baentsch, L. Baum, G. Molter, S. Rothkugel and P. Sturm. Enhancing the web's infrastructure: From caching to replication. IEEE Internet Computing. March 1997. P. 18-27.

M. Crovella and A. Bestavros. Self-similarity in world wide web traffic: Evidance and possible causes. In Proceedings of the 1996 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, May 1996. http://www.cs.bu.edu/faculty/crovella/papers.html

[DUS97] B. M. Duska, D. Marwood and M.J. Feeley. The measured access characteristics of WWW proxy caches. 1997. http://www.cs.ubc.ca/spider/marwood/Projects/SPA/Report/Report.html

[TAQ86] M. Taqqu and J. Levy. Using renewal processes to generate long-range dependence and high variability. In E. Eberlein and M. S. Taqqu, editors, Dependence in Probability and Statistics. pp. 73-89, Boston, 1986. Birkhauser.

[WES96] D. Wessels. Squid Internet Object cache. 1996. http://squid.nlanr.net/Squid/

[WIL97] W. Willinger, M. Taqqu, R. Sherman and D. Wilson. Self-Similarity through high-variability: Statistical analysis of ethernet LAN traffic at source level. IEEE/ACM Transactions on Networking, Vol. 5, No. 1, Febuary 1997.