Queensland University of Technology
Brisbane Australia

This is the author's version of a work that was submitted/accepted for publication in the following source:

Bhatia, Sajal, Mohay, George M., Schmidt, Desmond, & Tickle, Alan (2012) Modelling web-server Flash Events. In *2012 IEEE 11th International Symposium on Network Computing and Applications*, Conference Publishing Services, Cambridge, MA, USA, pp. 79-86.

This file was downloaded from: http://eprints.qut.edu.au/54456/

© Copyright 2012 IEEE

Notice: *Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source:*

http://dx.doi.org/10.1109/NCA.2012.24

Modelling Web-server Flash Events

Sajal Bhatia, George Mohay, Desmond Schmidt, Alan Tickle
Information Security Institute, Queensland University of Technology
GPO Box 2434, Brisbane 4001, Queensland, Australia
{s.bhatia, g.mohay, desmond.schmidt, ab.tickle}@qut.edu.au

***Abstract*—A Flash Event (FE) represents a period of time when a web-server experiences a dramatic increase in incoming traffic, either following a newsworthy event that has prompted users to locate and access it, or as a result of redirection from other popular web or social media sites. This usually leads to network congestion and Quality-of-Service (QoS) degradation. These events can be mistaken for Distributed Denial-of-Service (DDoS) attacks aimed at disrupting the server. Accurate detection of FEs and their distinction from DDoS attacks is important, since different actions need to be undertaken by network administrators in these two cases. However, lack of public domain FE datasets hinders research in this area. In this paper we present a detailed study of flash events and classify them into three broad categories. In addition, the paper describes FEs in terms of three key components: the volume of incoming traffic, the related source IP-addresses, and the resources being accessed. We present such a FE model with minimal parameters and use publicly available datasets to analyse and validate our proposed model. The model can be used to generate different types of FE traffic, closely approximating real-world scenarios, in order to facilitate research into distinguishing FEs from DDoS attacks.**

***Keywords*-Flash Events, Modelling, DDoS Attacks**

I. Introduction

The term *Flash Crowd* (FC) was coined by Larry Niven in a 1973 science-fiction story in which huge crowds of people went back in time to visit historic events [1]. In the World Wide Web context, a Flash Event (FE) refers to the situation in which a large number of legitimate users concurrently send access requests to a web-server, either following some newsworthy event or as a result of redirection from popular websites like Slashdot or other social media. This can lead to a dramatic increase in incoming traffic and hence QoS degradation. Some examples of FEs include popular news websites experiencing a surge in incoming traffic after major world events (earthquakes, 9/11 attacks), major sporting events (Olympics), or presidential elections.

A Denial-of-Service (DoS) attack is an explicit attempt by an attacker to disrupt the services of a provider intended for its legitimate clients by consuming computing and networking resources. DDoS is the distributed form of DoS, in which an array of compromised machines are used to attack a server. Efficient and accurate detection of DDoS attacks still remains an unsolved problem despite extensive research. Ever-changing attack vectors and their increasing complexity, as well as the potential for confusion with FEs, make it difficult to build a generic detection model.

The degradation caused by unexpectedly high levels of traffic, whether occasioned by DDoS or FE activity, can have economic consequences and requires remediation. Recent reports by Amazon [2] suggest that even a 100 ms delay in response time causes an approximately 1% drop in overall sales. Upon detection of DDoS, attack mitigation mechanisms must be activated to filter out malicious traffic and sustain uninterrupted services to genuine clients. Detection of a FE, on the other hand, may be followed by the activation of various load-sharing mechanisms such as Content Distribution Networks [3] to accommodate additional legitimate clients trying to access web resources. One of the prime differences between these two network anomalies is their source. DDoS attacks are often caused by an array of compromised machines, whereas a FE results from requests by legitimate clients. However, there have been cases where DDoS attacks have resulted from coordinated 'hacktivists' such as Anonymous instead of compromised and pre-programmed machines[1]. In any event, a detailed study of FEs and their characteristic features is essential in order to understand how they may be differentiated from similar looking DDoS attacks. We intend to use our results to generate synthetic FE datasets which can closely approximate a real-world scenario.

In this paper we attempt to address this issue by providing a comprehensive study of FEs, categorising them as Predictable, Unpredictable and Secondary. We characterize FEs in terms of three key components and present a server-side model with minimal parameters that captures various characteristic properties (duration, shape, intensification, source IPs, resources accessed) of different FEs. Finally, we present a detailed analysis of some publicly available datasets to validate our proposed model.

The remainder of the paper is structured as follows. Section II gives an overview of recent work done in the field of FE modelling and simulation. Section III provides a classification of FEs into three broad categories, while Section IV presents our proposed FE model based on three characteristic components. Section V presents a detailed analysis of some available datasets of FEs in order to

[1] http://www.computerworld.com/s/article/9218528/

validate the proposed model. Section VI summarizes the work and describes the future directions of our research.

II. Background and Related Work

A significant amount of research has been conducted in workload characterization of web-servers [4]. However, much of this work has been directed towards understanding typical web-traffic behaviour to improve caching and content distribution capabilities [3]. Workload analysis of web-servers during extremely heavy traffic situations, as in a FE, is comparatively less researched. Comprehensive modelling of FEs to describe their various aspects, and synthetic generation of realistic FE traffic, needs to be explored to better understand this network anomaly.

Ari et al. [5] proposed a simple model for Flash Crowd traffic consisting of three phases: a ramp-up phase, a sustained traffic phase and a ramp-down phase. Their proposed model focuses on the time duration of each of the phases (controlled by a parameter 'shock_level') and assumes a linear increase and decrease in the incoming traffic. This seems overly simplistic as behaviour is likely to vary significantly between different types of FEs e.g., the traffic seen during *Slashdot triggered FEs* tends to increase quickly by a large amount and then fade off slowly [6]. Such FEs occur when short news articles are posted along-with a Web link redirecting the reader to another web-server containing a full description of the associated story, and the receiving web-server can be quickly overwhelmed. In addition, their model only represents incoming traffic. Research presented in [7] analysed CoralCDN traffic, an open content distribution network, and defined flash crowds as successive intervals of time for which the rate of incoming requests to a fully-qualified domain name increases exponentially. However, that work focuses on aspects of online content service delivery, rather than on developing a detailed model of FEs.

Bodik et al. [8] analysed five datasets for volume and data spikes and used a closed-loop workload generator to synthesize workload traffic in a test environment. Their model is very similar to one proposed by [5] i.e. a linear ramp-up and a ramp-down of traffic during a FE. Zhang et al. [9] presented a model for the *magnitude* of FCs (flash crowds) in Bit Torrent. That model takes into account the fact that the service capacity of a Bit Torrent *swarm* (a group of peers sharing a torrent) increases with its peers, unlike FCs seen in web-servers with finite capacities.

Research [3] characterized FCs and DDoS attacks, however it was based on the analysis of some private datasets. There are only a limited number of public domain datasets, representing various types of FEs and most are web-server logs in Common Log Format, which makes it difficult to use them for experimentation, e.g. by replaying the traffic over a network to test various FE and DDoS detection mechanisms, or by measuring their effects (such as CPU, Memory and Bandwidth utilization) on the target server. Although this lack of representative datasets is somewhat limiting, there is nonetheless sufficient data available to make useful progress in developing models of FE traffic.

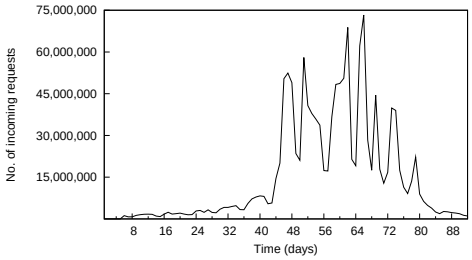

Figure 1. World Cup incoming traffic profile

III. Classifying Flash Events

The focus of our work is to model FEs by dividing them into three broad categories: *predictable*, *unpredictable* and *secondary*. Research presented in [3] briefly describes the predictable and unpredictable nature of FEs. We extend that work by presenting a detailed classification of FEs based on some real-world examples and make use of our classification subsequently in Section IV to propose a general FE model.

A. Predictable FE

A *predictable* FE (henceforth pFE), can be defined as one whose expected occurrence is known *a priori*, thus allowing network administrators to prepare for them using various provisioning techniques such as load-sharing mechanisms or CDNs. Some examples are product releases (e.g. by hi-tech companies like Apple) or widely followed sporting events such as the Olympics, where the expected time of the incoming traffic burst is well known in advance. The time when the incoming traffic will hit its peak can also be fairly accurately estimated. Most pFEs are directed against servers owned by big companies who can afford the necessary load or content-sharing techniques to mitigate their effect.

The 1998 FIFA World Cup dataset [10] and NASA web-server logs [11] are some of the few datasets available in the public domain representative of a pFE. Figure 1 shows the daily traffic volumes experienced by the World Cup websites. Each of the individual peaks or bursts of traffic is essentially fine structure within the overall period and represents individual FEs on a smaller scale than the overall event [4]. We have analysed the traffic during the two semi-final matches (73^{rd} and 74^{th} day) of the 1998 FIFA World Cup traffic (Figure 1) for this paper. Figure 2 presents the semi-final peak requests.

B. Unpredictable FE

Events that are totally unexpected can, if sufficiently newsworthy, cause a sudden and dramatic surge in network traffic to a site that is thought to describe the event or provide further leads. We use the term *unpredictable* FE (henceforth uFE) to describe the ensuing burst of network

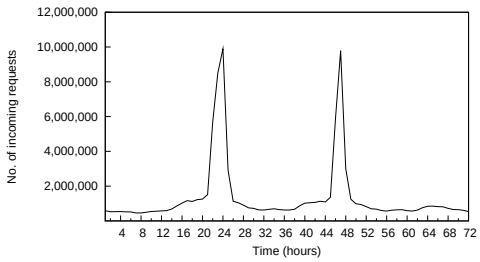

Figure 2. World Cup semi-finals traffic profile

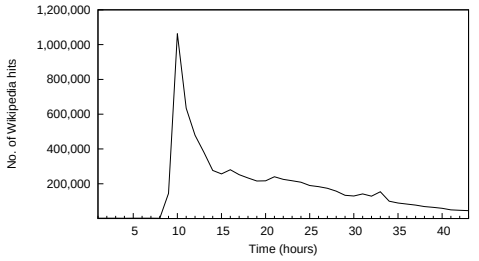

Figure 3. Hourly hits on Wikipedia after the death of Steve Jobs

traffic. Provisioning for these events in advance is akin to preparing for natural catastrophes like a Tsunami or an Earthquake. Designing systems to handle such a catastrophe is possible but may be economically infeasible due to its rarity. The 9/11 terror attack led to such an uFE when major news websites like CNN and MSNBC were overwhelmed by the amount of incoming traffic, pushing their availability close to 0% within minutes after its occurrence [12]. The start and peak-load time of such events is unpredictable and sometimes difficult to identify *post hoc*. Their occurrence frequency is relatively lower than pFEs. Figure 3 is an example of a uFE when the popular website Wikipedia experienced a sudden increase in its hourly hits following the death of Steve Jobs[2]. Similar traffic was also observed following the death of Michael Jackson.

C. Secondary FE

There is a third category of FE, which we call a *secondary* FE (henceforth sFE). These kinds of FE usually occur when a brief article, alongwith a web-link, often related to an interesting news item, but not as newsworthy on a world-wide scale as a uFE, is posted on widely followed websites like Slashdot. This can capture the attention of a large number of followers and redirect a high percentage of them to another website in search of additional information. When these (usually user-posted) articles contain links to poorly resourced websites, they can easily result in the redirection of an unprecedented amount of traffic to these small websites, which exceeds their available resources and eventually cripples them. Once again, the event (article posting) is unpredictable, and the peak-load time is likewise

[2]http://dom.as/2011/10/07/steve-jobs/

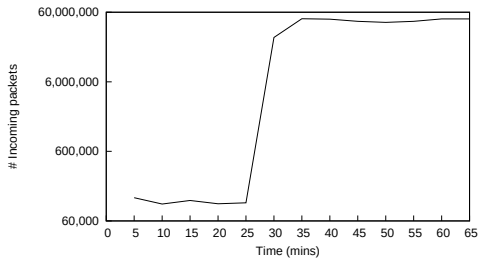

Figure 4. Incoming traffic profile for CAIDA DDoS attack dataset

relatively difficult to predict. Provisioning for such events can be challenging but is more feasible than for uFEs. Anderson [6] show the incoming traffic patterns of a sFE. The receiving server experiences a *'phase-transition'* from virtually no traffic to heavy traffic in a short time.

Table I summarizes the classification of FEs in terms of *Predicted Start-time*, *Predicted Peak-load Time*, *Typical Peak Value*, *Provisioning* and *Occurrence Frequency*. We argue also that most of pFEs are against web-servers which can afford the cost of web-content outsourcing though CDNs and other load sharing mechanisms. Hence, even during peak load they are not as badly affected as those subjected to sFEs and uFEs e.g., a Slashdotted website or various news websites after a large scale natural calamity.

IV. Flash Event Modelling

For maximum utility, our work is intended to model FEs using only available server-side information. We hypothesise that a FE can be described in terms of the following three components: volume of incoming traffic, source IP addresses generating the traffic volume and web-resources accessed during the FE. We aim in future work to use these components along with some other server-side system-based parameters like CPU utilization, to identify FEs and to distinguish them from DDoS attacks, which are inevitably similar in some respects. We have done some preliminary work on the rate of change of new IP addresses sourcing website traffic [13, 14], also on how that varies between DDoS and FE activities [15], and intend to consolidate and extend that work using the FE model developed here.

A. Volume of Incoming Traffic

In contrast to a DDoS attack, during a FE (more precisely during a sFE or uFE) it is unlikely that the entire web community gets simultaneously informed and goes online to get additional information. It usually takes time for the news that is the root cause of a FE to spread across the world. Hence, even though the incoming traffic to the web-server increases sharply, it is not as immediate as in a DDoS attack, assuming it does not mimic a FE. Figure 4 shows the rate of incoming traffic for a CAIDA 'DDoS Attack 2007 Dataset' [16]. Contrasting this with Figures 2 and 3 demonstrates that the rate of incoming traffic for a DDoS attack is much higher

Table I
FLASH EVENT CLASSIFICATION

FEs	Predictable Start-time	Predictable Peak-load Time	Typical Peak Value	Provisioning	Occurrence Frequency
pFE	Yes	Yes	Medium	Feasible	High
uFE	No	No	High	Not feasible	Low
sFE	No	No	Medium	Not Feasible	Moderate

than for a FE (note that whereas Figure 4 shows Time in minutes, Figures 2 and 3 show it in hours).

For our traffic model, we define the ratio of the 'peak request rate' (R_p) to the 'average normal request rate' (R_n), as observed by a web-server during a FE, as a 'traffic amplification factor' (A_f) (similar to the *shock_level* parameter of [5]). Each of the request rates R_n and R_p are expressed as the number of incoming requests per unit time. Thus:

$$A_f = \frac{R_p}{R_n}$$

A_f can be set to different values in our model to simulate various FE scenarios. We consider a FE to have two major phases: a *flash-phase* and a *decay-phase*, and argue that the 'plateau period' or *sustained traffic phase* [5, 8] is extremely short lived and orders of magnitudes smaller than that of the other two phases, the flash-phase and the decay-phase. Our analysis of real-world datasets in Section V shows the absence of any such plateau period.

The duration of the flash-phase and the decay-phase are represented by ΔT_f and ΔT_d respectively. ΔT_f represents the time it takes for the normal request rate (R_n) to reach the maximum i.e. R_p. ΔT_d is the time taken for the peak request rate to decay to the normal request rate (R_n). We argue that there is a difference in the relationship between the durations of these two phases for the three different types of FEs and first discuss this in regard to pFE. In the case of a football match (and indeed many sporting events), the event has a duration, from the start through to its completion, and while people may be interested in checking throughout, nonetheless we can arguably regard completion of the event and posting of the final score as 'the event of interest'. At that point we expect a peak in the traffic volume and the rate of incoming requests to the web-server to quickly decline and return to 'normal'. If the web-server has difficulty in handling the peak volume of requests, then persistent delays (response time) from the server side owing to network congestion, may force users to secondary sources of information, thus causing a still faster decline in the incoming traffic rate. For these reasons we posit that for a pFE the decay time is comparable and possibly less than the build-up time.

$$\Delta T_f \geqslant \Delta T_d$$

This differs from the research presented in [5] which models ramp-down duration (ΔT_d in our case) as 'n' (constant) times the sustained-phase duration.

However, in the case of uFEs and sFEs, there is no such time-based 'end of event'. Long after the actual occurrence of the event, some section of the web-user community will likely still be interested in it. Thus, there is a slow and gradual decay in the incoming request rate for sFEs and uFEs as compared to pFEs. Therefore, we can say that the *flash-phase* lasts for much smaller duration than the *decay-phase* in sFEs and uFEs or $\Delta T_f << \Delta T_d$ We now discuss the two phases of a FE.

Flash-phase: During a FE, the excess load on the server is mainly due to an increase in the overall number of clients accessing the web-resource rather than in the number of requests per client [3]. Apart from a small percentage of enthusiastic clients, a majority of the clients participating in a FE are mostly interested in a very specific set of information related to that event [3]. Thus, we would expect that the average number of requests per participating client would remain relatively constant during a FE.

When a newsworthy event occurs, it generally takes finite time for the news related to the event to spread across the world. More and more people get interested and go online to investigate. This leads to a dramatic increase in the interested user population, which can be represented in terms of the classic exponential growth model used in many other domains. The exponential growth in user population appears intuitive for uFEs more so than for pFEs. However, we tested our proposed model against all three types of FE.

Based on these two propositions (relatively constant per client request rate and exponential growth rate of interested population), we can say that during the *flash-phase* the increase in rate of incoming requests is proportional to the current rate i.e. it varies with time (t) as:

$$\frac{dR}{dt} = \alpha R$$

where α is a *flash-constant*. Assuming that the *flash-phase* starts at time $t = t_o$ where $R = R_n$, we obtain:

$$R = R_n e^{\alpha(t-t_o)}$$

The end of the *flash-phase* is marked by time $t = t_f$ where the incoming request rate reaches the peak i.e. $R = R_p = A_f \times R_n$ Substituting these values gives the *flash-constant*:

$$\alpha = \frac{\ln A_f}{(t_f - t_o)}$$

and the value for the incoming traffic during the *flash-phase* (i.e. for $t_o < t \leqslant t_f$)

$$\boxed{R = R_n e^{\frac{\ln A_f}{(t_f - t_o)}(t-t_o)}} \quad (1)$$

The start of the flash-phase i.e. t_o and the time of the peak request rate (R_p) can be determined *post hoc* visually or in real-time using Change Point Analysis (CPA). The exponential increase of incoming traffic during the flash-phase differs from the research presented in [5, 8] which models the ramp-up phase as a linear function, and [7] defines flash crowds using a quadratic growth model.

Decay-phase: At the start of this phase, the number of incoming requests is at its peak and starts to decline. We speculate three possible reasons for this behaviour. Firstly, by this time the most interested users have the information they were looking for and thus have moved on. Secondly, the main web-server hosting the information reaches its serving capacity and thus starts rejecting new connections. And lastly, a prolonged response time from the primary web-server starts to annoy users, forcing them to either return later or to look for other sources of information (secondary servers). The last speculation intuitively suggests that a growth, possibly exponential, in the number of secondary servers, providing the same information as the primary server, contributes to the exponential decay in the number of incoming requests to the main web-server. Hence, we argue that during the *decay-phase*, the rate of incoming requests (R) decreases with time (t) as follows:

$$\frac{dR}{dt} = -\beta R$$

where β is a *decay-constant*. At $t = t_f, R = R_p$ and at $t = t_d, R = R_n$, thus

$$\beta = \frac{\ln A_f}{(t_d - t_f)}$$

and, the incoming traffic model during the *decay-phase* i.e. for $t_f < t \leqslant t_d$

$$\boxed{R = R_p e^{\frac{-\ln A_f}{(t_d - t_f)}(t - t_f)}} \quad (2)$$

where $R_p = A_f \times R_n$. The exponential decay model of incoming traffic during the decay-phase (Equation 2) differs from [5, 8], which show a linear ramp-down and [7], which models only the ramp-up phase. Before the *flash-phase* ($0 < t \leqslant t_o$) and after the *decay-phase* ($t > t_d$) we have:

$$\boxed{R = R_n} \quad (3)$$

In summary, we have modelled the traffic volume for each of the two main phases of a FE i.e. *flash-phase* and *decay-phase*. We represent a FE as a set of three equations (1, 2 and 3) with few configurable parameters (R_n, R_p, t_o, t_f and t_d) which are tuned to represent different FEs.

B. Source IP Addresses

The dramatic increase in the number of incoming requests during a FE can be attributed to either a substantial increase in the number of requests per participating client or an increase in the overall population interested in that particular event, or a combination of the two. Based on two FE datasets, research [3] concludes that the per-client request rate does not increase, but rather drops and remains lower during FEs compared to other times. The analysis of a pFE dataset presented later in this paper shows somewhat different results. It shows a slight increase in the number of requests per client with the onset of a FE. But it also shows that the overall increase in incoming traffic volume is mainly due to an increase in the interested population.

The analysis of a publicly available dataset presented in Section V shows how the number of source IPs increases with the onset of a FE (i.e. at $t = t_0$) and continues to increase during the *flash-phase* before starting to decline over the *decay-phase*. Thus, during the course of a FE, variations in the number of source IPs would be expected to closely resemble variations in the incoming traffic volume. Nonetheless, there are some reasons why the same model may not necessarily be applicable to source IPs. For instance, for source IPs, there are some complicating factors. Firstly, the number of requests per client may not be constant in all cases, and secondly, IP mapping (NAT, DHCP) may have the effect of invalidating the assumption that the number of clients is equivalent to the number of source IPs, thus affecting the assumption of constant packets per IP.

C. Resources Accessed

The third component we have used to understand and model FEs is the 'randomness of resources' being accessed. We speculate that during a FE, barring a small percentage of enthusiastic clients, most are normally interested in a very specific set of information relating to that event e.g. results of a match or presidential elections. In other words, this translates into a greater concentration of resources being accessed during the event as compared to the pre-event or post-event period.

The analysis of a publicly available dataset used in this paper confirms our speculation. The number of different resources being accessed during the course of a FE decreases considerably compared to other times. In this paper, we have used entropy as the metric to measure this component. The resource entropy starts to decrease, from its average value with the onset of the FE and continues to drop for the rest of the *flash-phase*, when most of the interested population start looking for a very specific set of resources. Once the event comes to an end, marked by the start of *decay-phase*, the randomness of resources being accessed starts to increase until the end of this phase, before returning to normal. Section V presents the analysis of this component.

V. Analysing Flash Events

Legal and privacy issues in obtaining real-world datasets means that only a very limited number of datasets are publicly available. Although some of these are old, they represent real FEs, mostly pFEs, and still exhibit most (if not

Table II
DATASETS USED

Dataset	Duration (hours)	Granularity	# Packets	# Unique Source IPs	# Unique Resources	FE Type
Semi-final 1	20	seconds	42,373,729	160,704	38,892	pFE
Semi-final 2	20	seconds	34,131,050	152,020	38,102	pFE
Wikipedia Hits	43	hours	7,417,070	N/A	N/A	uFE
Slashdotted Website	80	hours	N/A	N/A	N/A	sFE

all) of the key characteristics that might be used to define a FE. We have used traffic volume, source IP addresses and resources accessed, where available, to validate our model.

A. Datasets

In this section of the paper, we validate the three components (incoming traffic volume, source IP addresses generating the traffic, and web-resources accessed during the FE) of our proposed model by analysing some of the existing datasets available in the public domain, that represent different types of FE. The datasets used in the validation are briefly described below and summarized in Table II.

1998 FIFA World Cup (pFE): This dataset is provided by the Internet Traffic Archive [10] and contains all the HTTP GET requests (along-with the resources accessed) made to the 33 web-servers for the duration of the World Cup. The dataset is in Common Log Format with source IP addresses replaced by unique integer identifiers for privacy reasons. We analysed the two peaks in the incoming traffic corresponding to the semi-finals as two separate pFEs (see Figures 1 and 2) and used that analysis to validate our model. The dataset was sampled at one-minute intervals.

Hourly Hits on Wikipedia (uFE): This data was provided by Domas Mituzas, who maintains a repository of page view statistics for various Wikimedia projects[3]. We selected records following Steve Jobs' death as representative of a uFE. The data provides hourly aggregates of requests and covers a period of 43 hours (starting 16:00 hours UTC, 5^{th} October 2011) as shown in Figure 3.

Slashdotted Website (sFE): Since there were no public domain datasets representing a sFE, we used high-level statistics of empirical data [6]. Although incomplete, this shows network activity on web-servers in the Department of Geological Sciences at Southern Methodist University, after their website was listed on the front page of Slashdot on June 15, 2003 and August 10, 2003.

Table II shows that, except for the 1998 FIFA World Cup data, the other datasets available in the public domain lack both 'source IPs' and 'accessed resources'.

B. Analysis

We now present our analysis of the above datasets in terms of three components we have used to model a FE i.e. incoming traffic volume, source IP addresses generating that traffic and web-resources being accessed during the FE.

[3]http://dumps.wikimedia.org/other/pagecounts-raw/

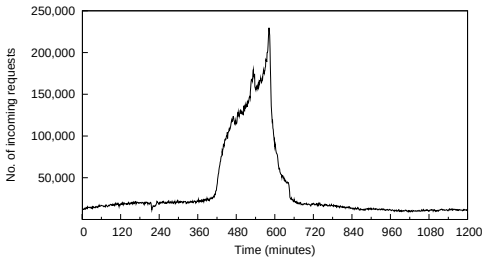

Figure 5. Incoming traffic during 1^{st} Semi-final

Incoming Traffic Volume: Figure 5 shows the actual incoming traffic volume over the 20-hour period around the 1^{st} semi-final match during the 1998 FIFA World Cup. The incoming traffic volume is represented as the number of packets received per one minute sampling interval. The rate of incoming packets before and after the FE is roughly constant. This feature is represented by equation (3) of our model. The match starts around the 420 minute mark on the x-axis (t_o), continues for the next 162 minutes (ΔT_f), before ending around the 580 minute mark (t_d). The 1^{st} semi-final match extended into a penalty shoot-out phase. During the *flash-phase*, R_n increased by a factor of nearly twelve.

Except for the (FIFA World Cup dataset), the other datasets were only available with hourly time resolution. Hence, for comparison purposes, we have used one-hour as the sampling interval for the FIFA World Cup dataset to find the values of the model parameters (R_n, R_p, t_o, t_f, t_d) as shown in Table III. The time-based parameters (t_o, t_f, and t_d) indicate the instances of time (in hours) when different phases started and ended and the requests rates (R_n and R_p) were calculated (for pFE) and observed (for uFE and sFE) on a per-hour basis for comparative analysis. Table III also shows the values for derived parameters viz. the incoming Traffic Amplification Factor (A_f), α (flash-constant), β (decay-constant), and the duration of flash and decay phase i.e. ΔT_f and ΔT_d, for the three datasets used.

It is interesting to note that the incoming traffic amplifies (A_f) by nearly 600 times in case of the uFE. This substantial increase in incoming traffic, combined with its relative infrequency, makes it less feasible to have any provisioning for such events. Another observation is the rate at which the incoming traffic increases and decreases for each FE as determined by α (flash-constant) and β (decay-constant) respectively. These constants have comparable values for the two pFEs, whereas in sFE and uFE they differ substantially. In the latter cases the incoming traffic tends to fade-off at

Table III
FLASH EVENT TRAFFIC MODEL PARAMETERS

FEs	$\mathbf{R_n}$	$\mathbf{R_p}$	$\mathbf{t_o}$	$\mathbf{t_f}$	$\mathbf{t_d}$	$\mathbf{A_f}$	$\mathbf{\Delta T_f}$	$\mathbf{\Delta T_d}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$
pFE: 1998 FIFA World Cup (1^{st} Semi-final)	1,167,621	9,898,273	7	10	12	8.47	3	2	0.71	1.06
pFE: 1998 FIFA World Cup (2^{nd} Semi-final)	1,032,715	9,785,128	8	10	12	9.47	2	2	1.12	1.12
uFE: Hourly hits on Wikipedia	1,826	1,063,665	8	10	43	582.51	2	33	3.18	0.19
sFE: Slashdotted Website	1,000	78,000	13	15	71	78.00	2	56	2.17	0.07

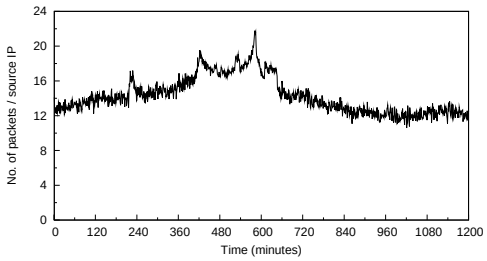

Figure 6. Packets per Source IP during 1^{st} Semi-final

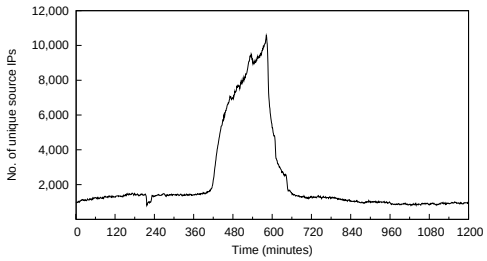

Figure 7. Unique source IPs during 1^{st} Semi-final

a much slower pace as compared to the build-up phase, whereas for the pFEs the incoming traffic appears to increase and decrease at similar speeds. These parameters can be configured and we intend to use them subsequently to generate different types of FEs.

Source IP Addresses: As described previously, we expect the per-client request rate to be almost constant during the event. The analysis of the 1^{st} semi-final match dataset, however, slightly deviates from this speculation. It instead shows a small increase in the per-client request rate for the duration of the FE as shown in Figure 6, which plots the number of requests per unique source IP against time for the 1^{st} semi-final. Thus, it is the overall increase in the participating population that causes a surge in the incoming traffic during a FE. Figure 7 shows the rate of increase and decrease of unique source IPs during the event. It varies in a similar fashion to the incoming traffic (Figure 5).

We conclude that during a FE, the majority of the traffic increase is due to an increase in the overall participating population, although there is also a slight increase in the per-client request rate as compared to the non-flash period. This result differs from [3], which states that the per-client request rate drops and remains lower during the flash period as compared to other time periods.

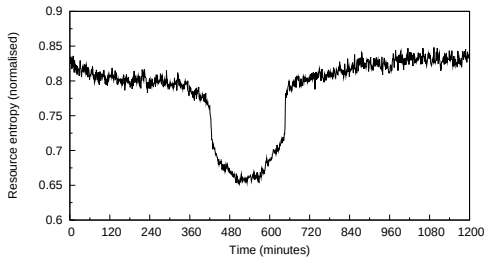

Figure 8. Resources accessed during 1^{st} Semi-final

Accessed Resources: As the news causing the FE spreads across the Web, more and more people go online to get the related information. We argue that a majority of the participating population is generally interested in obtaining a specific set of information from the web-server. Thus the number of different resources being accessed by the client during a FE should be lower than during non-flash-event times. One way to measure this attribute is Shannon entropy, a measure of the uncertainty associated with a random variable. Shannon's formula [17] is used to compute Entropy H, where $0 \leqslant H \leqslant log_2 N$, and N is the number of discrete random variables X. In our analysis X represents a *unique web-resource* being accessed. We have used normalised entropy H_o, where $0 \leqslant H_o \leqslant 1$, given by:

$$H_o = -(\sum_{i=1}^{N} p_i log_2 p_i)/log_2 N$$

where p_i is the probability of that random variable. In our analysis, the probability p_i of each resource is its relative frequency of occurrence in each one minute interval. Experiments conducted for one-second and five-minutes intervals gave similar results. Figure 8 shows a drop in the normalised resource entropy from the onset of the event and stays low for the entire duration of the event.

C. Model Validation

In this section we use the equations presented in Section IV, and values computed from real datasets, to generate synthetic data to see how closely the model can replicate real-world data. Due to the absence of resource request information, our validation was limited to the incoming traffic volume component of FEs. Also since, the dataset for sFE and uFE had hours as the granularity level, we decided to use the same granularity also for the pFE dataset. The 'belly' shape in the original pFE data (between hours

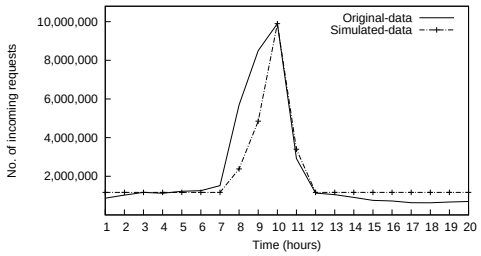

Figure 9. Traffic comparison for 1^{st} Semi-final

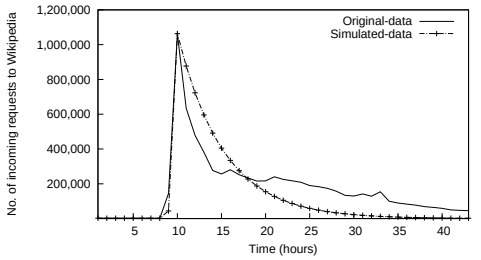

Figure 10. Traffic comparison for Wikipedia hits

7 and 10) seems to reflect a faster initial increase in the traffic than our model followed by a tailing off towards the peak. We intend to study this further in future work. Figures 9 and 10 compare real and synthetic data, for pFE and uFE, generated using the proposed FE model for incoming traffic. It shows that a simple exponential model, with few configurable parameters, can be used to capture the essential characteristics of a FE. This can be used to generate realistic FE traffic, thereby facilitating research in FE detection and its differentiation from DDoS attacks. Similar results were obtained for sFE but as the original data was unavailable [6], we opted not to include the results for synthetic data.

In contrast, the linear model postulates a plateau phase not found in our analysis and a linear ramp-down phase, which less closely models the actual data. The quadratic model deals only with the ramp-up phase (cf. sections II and IV).

VI. Conclusion and Future Work

We have shown that FEs can be closely approximated by a simple mathematical model consisting of an exponentially increasing flash phase and a decreasing decay phase, regulated by an amplification factor. We have further classified FEs into predictable, unpredictable and secondary. The particular traffic profiles of each of these events can be accurately modelled by varying a small set of configurable parameters including the amplification factor and the time durations of the flash and decay phases. Our future work will be directed toward applying this model to the synthetic generation of FEs with a view to developing techniques for successfully distinguishing FEs from DDoS attacks.

This research was supported in part by the Australia-India Strategic Research Fund.

References

[1] L. Niven, *The flight of the horse*. Ballantine Books, 1973.
[2] G. Linden, "Make data useful," *Presentation, Amazon, November*, 2006.
[3] J. Jung, B. Krishnamurthy, and M. Rabinovich, "Flash crowds and denial of service attacks: Characterization and implications for CDNs and web sites," in *Proceedings of the 11th International Conference on World Wide Web*. ACM, 2002, pp. 293–304.
[4] M. Arlitt and C. Williamson, "Web server workload characterization: The search for invariants," in *Proceedings of ACM SIGMETRICS Performance Evaluation Review*, vol. 24, no. 1. ACM, 1996, pp. 126–137.
[5] I. Ari, B. Hong, E. Miller, S. Brandt, and D. Long, "Managing flash crowds on the internet," in *Proceedings of 11th IEEE/ACM International Symposium on Modeling, Analysis and Simulation of Computer Telecommunications Systems (MASCOTS)*. IEEE, 2003, pp. 246–249.
[6] D. P. Anderson, "Surviving slashdot'ing with a small server," http://www.geology.smu.edu/~dpa-www/attention_span/, 2005, [Online; accessed 9-June-2012].
[7] P. Wendell and M. Freedman, "Going viral: flash crowds in an open CDN," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 549–558.
[8] P. Bodik, A. Fox, M. Franklin, M. Jordan, and D. Patterson, "Characterizing, modeling, and generating workload spikes for stateful services," in *Proceedings of the 1st ACM symposium on Cloud computing*. ACM, 2010, pp. 241–252.
[9] B. Zhang, A. Iosup, J. Pouwelse, and D. Epema, "Identifying, analyzing, and modeling flashcrowds in bittorrent," in *Proceedings of Peer-to-Peer Computing (P2P), 2011 IEEE International Conference on*. IEEE, 2011, pp. 240–249.
[10] M. Arlitt and T. Jin, "1998 world cup web site access logs," http://www.acm.org/sigcomm/ITA/, 1998, [Online; accessed 9-June-2012].
[11] J. Dumoulin, M. Arlitt, and C. Williamson, "Nasa web server logs," http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, 1995, [Online; accessed 9-June-2012].
[12] J. Hu and G. Sandoval, "Web acts as hub for info on attacks," *Retrieved April*, vol. 26, p. 2003, 2001.
[13] E. Ahmed, G. Mohay, A. Tickle, and S. Bhatia, "Use of ip addresses for high rate flooding attack detection," *Security and Privacy–Silver Linings in the Cloud*, pp. 124–135, 2010.
[14] G. Mohay, E. Ahmed, S. Bhatia, A. Nadarajan, B. Ravindran, A. B. Tickle, and R. Vijayasarathy, "Detection and mitigation of high-rate flooding attacks," in *An Investigation into the Detection and Mitigation of Denial of Service (DoS) Attacks*, S. Raghavan and E. Dawson, Eds. Springer, 2011, pp. 131–181.
[15] S. Bhatia, G. Mohay, A. Tickle, and E. Ahmed, "Parametric differences between a real-world distributed denial-of-service attack and a flash event," in *Proceedings of Sixth International Conference on Availability, Reliability and Security (ARES), 2011*. IEEE, 2011, pp. 210–217.
[16] P. Hick, E. Aben, K. Claffy, and J. Polterock, "The CAIDA "DDoS Attack 2007" Dataset," http://www.caida.org/data/passive/ddos-20070804_dataset.xml, [Online; accessed 9-June-2012].
[17] C. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.