Flash Crowd Detection within the realms of an Internet Service Provider (ISP)

Angelos Marnerides, Dimitrios P. Pezaros and David Hutchison

Computing Department
Infolab21
Lancaster University
Lancaster, LA1 4WA UK
{a.marnerides, dp, dh}@comp.lancs.ac.uk

Abstract- **It is truly a challenge to detect a network phenomenon with an unpredictable persona. Due to the simultaneous dependency on network traffic and end users, events such as Flash Crowds are hard to predict. This paper introduces a Flash Crowd (FC) prediction methodology to operate at the edges of the ISP network hosting the hot-spot (i.e. end-server). Such a methodology would act as the logic unit in the detection architecture that we also propose. The proposed methodology promotes prediction with the use of a mathematical relationship between the request and response rate subject to the assumption that a FC is composed as a linear state model.**

Keywords- **flash crowd, flash event, network traffic, anomaly detection**

I. INTRODUCTION

The term ***Flash Crowd*** was first baptized in 1971 in a science fiction story [1] where a Flash Crowd in that story was the situation when thousands of people were trying to go back in time to view historical events [2]. Similar behavior is observed for web and P2P objects when a large number of clients simultaneously access a specific data object offered by a particular server. With the use of networking concepts we define a Flash Crowd (FC) or Flash Event as ***the phenomenon where a server becomes unable for a period of time to respond to all the numerous requests produced by various clients in a limited amount of time within a network.*** There is evidence that such unexpected events have compromised major news-web cast sites such as MSNBC [3, 4], as well as unpopular sites that got virtually famous after being mentioned in a popular news feed [4]. In this work we are mostly concerned with the latter case which is also known as the ***Slashdot Effect*** [5].

The effects of such an event cause dramatic changes not only on the web server but also within the actual network that suffers from overloading. The latter is considered by us as our main axis for supporting that a network should not be used as a "black-box" component in the overall process of detection and mitigation of a FC. Overloading is created due to the large number of requests traversing the links within a network and in parallel due to the large number of responses where those contain numbers of larger packets in size and inject more load than requests do, resulting to cause transmission latency and congestion [11].

It is also important to mention that FCs promote a challenge to P2P architectures where even though known for their improved scalability and techniques to remediate bursty requests (i.e. distribution of file-chunks among peers) compared with the simple client/server model, still suffer from such events. A FC on an overlay affects the traffic exchange taking place on the underlay (i.e. an ISP). However, this paper is not concerned with the behavior and detection of FCs on an overlay scenario but instead proposes a detection mechanism on underlay scenery based upon the characteristics of a FC within the realms of a single ISP. We take the underlay – ISP scenario because we believe that current methods of over-provisioning applied by network operators do not facilitate properties for efficient congestion management and in parallel a correct confrontment for unpredictable events like a FC. No operator can guess the magnitude that an FC might export in a network and not every single-link can be massively over-provisioned just in case of a FC. Thus, the need for developers and maintainers to assess memory/capacity capabilities that a server should support in order to confront a FC is relaxed, since the phenomenon is detected before it takes a full and large scale at the server's network access link.

Our detection methodology is novel in that it focuses on the relationship between the request and response rate on the border routers of an ISP. This relationship is considered as the input to our three-step change detection mechanism for predicting the flash crowd signature even before the FC is committed on the web server. The prediction and further detection itself is based on a simple mathematical relationship of these two factors that their values are gathered on the initial ramp-up phase of the FC.

Our three-step prediction methodology is composed by having as its foundations the properties of the two linear state models that we present. The former, a linear state representation of the FC request rate is the result of work done in [3 ,12] as for the latter is a model that we introduce and denotes the linear state of the FC response rate. In order to predict correctly and avoid false alarms we define a range of observational points based on these two models and compare in theory whether their linear properties are still true. The novelty appears on the actual metric taken in the comparison of the linear properties which is basically the slope of the linear shape determining each increase of requests or decrease in responses. We propose that this slope is the most vital factor for detecting a change in traffic where it might expose a phenomenon such as a FC. In contrast with work done in [3, 12] or even [11, 13] our work does not consider the network as a black-box and it is not assuming that our mechanism is close to the hot-spot within an ISP. Nevertheless, it is undoubtedly acknowledged that FC classification through modeling and further detection was a topic studied by several people in the research community and this is why this paper dedicates the next section on some significant to mention achievements.

This paper is organized as follows. In Section II we discuss related work in the area of network traffic anomalies classification and FC detection. Section III introduces our overall FC detection approach. There is an explanation of the FC traffic-related linear generic models and in addition there is a detailed reference on our prediction methodology that composes the logic unit of our overall detection architecture. We discuss about future work in Section IV and finally Section V concludes the paper.

II. Related work

Within the work in the area of traffic anomaly detection there are three basic approaches for studying features and characteristics coming from a packet, flow and aggregate-level perspective.

For instance, in [6] after gathering a 6—month range data of SNMP and IP flows from a border router on a single Autonomous System (AS), researchers used wavelet filters in order to expose detailed characteristics of both ambient and anomalous traffic. Each dataset was categorized based on its signal frequency and using a pseudo-spline filter tuned on certain aggregate levels. The pseudo-spline filter was used in order to set certain recognizable frequency frames to the data collected. The organized set of data processed by the pseudo-spline filter was called strata where that was a hierarchy of component signals shown on aggregates. After that filtering it was then feasible to expose specific and exact characteristics for each phenomenon. Although this work was significant in terms of traffic categorization, still there was a lack of an on-line/runtime detection-prediction mechanism to deal with the unpredictable character that a FC might trigger.

This prompts to [7] where there was a combination of statistical measurements and filtering in order to compose a classification and further runtime detection mechanism for each phenomenon. As a basis was the calculation of the traffic matrix within the network by taking measurements of origin-destination flows as this was done in a previous work [8]. This process came after the assumption of denoting the network as a linear state space model. The traffic matrix calculation was one of the factors determining the possible evolution of the network and with the help of a Kalman filter [9] there was the capability to filter out normal traffic and extract the anomalous behaviour. Again, via this work there was not a direct emphasis on the exact FC behaviour and characteristics.

An alternative approach on identifying and then mitigating FCs was the approach taken by Jung et. al.[2]. There was a suggestion for an adaptive Content Distribution Network (CDN) architecture, which was indeed useful in manners of cooperative caching, and the dynamic delegation technique proposed improved a fair distribution of client requests within selected components in their architecture that helped the mitigation of the FC targeting the end server. A significant part of this work was the exact definition of the differences between a FC and a DoS attack, two distinct phenomena which are manifested by a set of very similar (close to identical) characteristics from a first point of view and analysis on a simple tool such as NetFlow [10].That approach was considering mainly metrics taken on the end host rather than the actual traffic in the network [6].

A vital reference point to our work is the Network Early Warning System (NEWS) developed and evaluated with success in [11]. NEWS is a router-based system performing FC detection using the web-response performance metric and then mitigating the phenomenon using an aggregate-based control between requests and responses. Its main FC detection component (which was the issue of interest for us) is purely based on the decremental behavior of the number of responses sent from the web-server. Request rate was used for FC mitigation, rather than detection. This was the main reason for us to introduce a detection methodology that combines the two factors of request increase and response decrease and compose a unique factor that determines the FC detection on a network-level approach.

III. FC Detection

We have considered as foundations for our methodology two general models that we present in this section. The first represents the behavior of the average request rate in respect with the overall time frame where an FC is created and the second presents the characteristics of the average response rate, again within the same time range. The former has already been defined in [4, 13] whereas the latter is something that never came in to our knowledge in such an abstract and generic form. This generic model was developed according to traces from real FC incidents documented in the literature [11, 13].

A. *FC Request Rate Model*

According to [4, 12] a FC is decomposed in to three phases each of which states the relationship between the requests (which can be referred to in a generalized form as traffic) arrived on a server within a certain time interval. In our detection we assume (as in [11, 13]) that the request rate arriving on the server also appears on the border router of the local ISP where the end-server belongs to. In case of multihoming (i.e. two or more border routers) the number of requests is a simple addition of the rates experienced at the edges routers. This addition employs the composition of the request rate state for the particular time of measurement t that we define as Ψt .

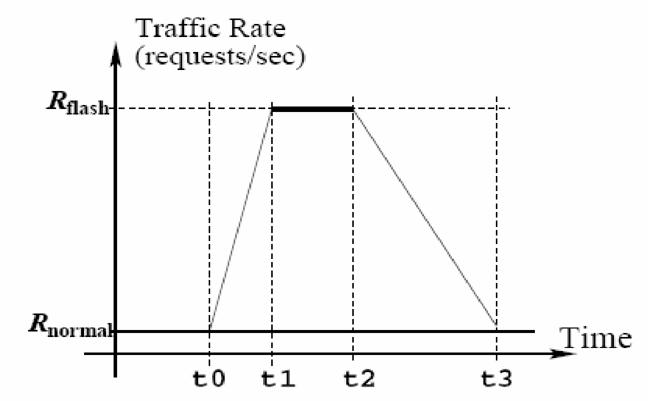

Figure 1: The FC Request Rate Model

Figure 1 shows these three phases in the event. Initially, there is the *ramp-up phase* that has a (theoretically) linear increase from a normal request rate *Rnormal* to the maximum rate of requests *Rflash* in a small amount of time *[t_0, t_1]*. Subsequently, there is the *sustained traffic phase* at *Rflash* between t_1 and t_2. Then, the traffic decreases and eventually returns back to *Rnormal* in the time interval between t_2 and t_3; this is referred to as the *ramp-down phase*.

Based upon this model we declare as the general FC request rate state Wt as:

$$Wt=\sum_{t=t0}^{t=t3}\Psi t \qquad (1)$$

B. *FC Response Rate Model*

We have extracted and modelled the state that an FC holds in respect with the response rate characteristics that the hot-spot promotes. Even though our resulting model denotes the abstract characteristics of the response rate behaviour it enabled us to go a step further and satisfy our assumptions for the linear state characteristics in a FC. In parallel there was a justification on the assumption that responses compose an opposite behaviour in comparison with the requests that a server accepts and that at the same time there is a direct dependency between them. Modelling and extraction was accomplished due to the comparison and analysis of the request rate presented in [11, 13] and indeed as it was assumed the response rate has a direct, depended and completely opposite relationship with the response rate during the relevant time intervals. A clearer view for this argument is shown in figure 2.

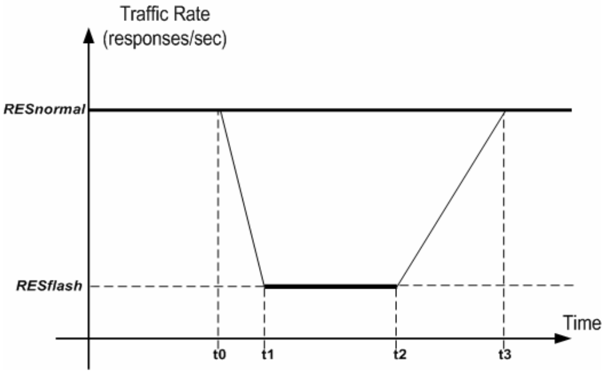

Figure 2: The FC Response Rate Model

Since the request rate (presented in the previous section) starts with a normal behaviour at *Rnormal* it was reasonable for us to assume that the same takes place on the response rate. Our assumption was justified with the analysis of the response rate measurements presented in [11, 13]. Therefore we define as the response rate normal behaviour at *RESnormal* as shown on the diagram in figure 2. Questionable was also the behaviour of responses while the requests ramp-up phase is taking place. This question was again answered by the experimental results in [11, 13] and from that point we were in the position to also identify the constant response rate while the FC occurs. Hence, within the time interval *[t0-t1]* is what we define as the responses ramp-down phase which is then followed by the interval *[t1-t2]* where this denotes the actual FC interval having the response rate reaching its minimum value *RESflash*. Finally, the interval *[t2-t3]* shows that the FC starts loosing its high-peak period in such manners that enables the server to start responding to requests up to *t3* where it reaches again its normal response behaviour up to *RESnormal*.

Here we follow the assumption that the response rate sent from the server arrives on the border router of the local ISP. This assumption is made to help us employ our detection before the sudden burst of requests reaches its highest peak on the server's access link. Furthermore, with the collection of information from that particular border(s) we are able to satisfy our following equation.

$$\Omega t=\sum_{t=t0}^{t=t3}Vt \qquad (2)$$

We define Ωt as the FC overall response rate state that is composed by the summation of all the observational response rate states within the time interval [t0-t3]. As equation 2 shows, each observational response rate state at any time t is defined as Vt. We support that with the knowledge of Vt at any time our detection mechanism is able to detect the change taking place in the response traffic and start predicting if there is a possibility for a FC. In addition, Vt estimation on the border router(s) will also set a pre-knowledge capability in our mechanism and be in the position to determine if the traffic passing by would be harmful and trigger congestion traffic within the ISP as according to [13] is not the burst of requests that causes congestion rather than the responses that hold larger capacity

C. *Prediction Methodology*

The prediction methodology composed was a result of confronting the specific phenomenon in an applied mathematics fashion. By setting as a foundation to our approach the two models specified in the previous sections, we were able to apply some basic concepts of space linear properties and regulate their characteristics in respect with the requirements for a FC prediction mechanism. Our approach was mainly directed towards the initial ramp-up and ramp-down phase coming from the request and response rate respectively. The main objective was to relate in mathematical terms under the spectrum of linear properties the observations taking place in the both phases.

As one of the main goals for us was to relate the observational request rate state Ψt with the observation we commit on the response rate state that we have already represented as Vt. In both cases and we also consider as a single observational point in the space determined by the request rate state space model as *(td, Rd)* and its subsequent related single observational point for the response rate space model as *(td, RESd).*

Therefore based on linear properties we define Ψt and Vt with the two following equations.

$$\Psi t = m(t - Rd) + td \qquad \textbf{(3)}$$

$$Vt = \lambda(t - RESd) + td \qquad \textbf{(4)}$$

Where in equation (3) property *t* denoted as the full range of time that any observation can take place, *Rd* as the request rate on time *td* and *m* as the changing factor of the state. Similarly in equation (4) property *RESd* denoted as the response rate on time *td* within observational range of time *t* with a changing observational factor for state *Vt* represented by λ.

The prediction mechanism we propose sets as its basis for rate change detection and FC classification the behaviour of variables *m* and λ within the range of time *[t0-t1]*. As already mentioned this range represents in both models presented earlier the ramp-up and ramp-down phase for requests and responses respectively.

As figure 3 shows, within the range of *[t0-t1]* both models represent their behaviour with a straight line. The request rate model presents a linearly increasing shape determined by a slope denoted by the tangent of the angle *k* (tan(*k*)). Similarly, the response rate model within the range *[t0-t1]* is represented by a decreasing linear shape where its state is depended upon a slope where we name as tangent of angle *b* (tan(*b*)).

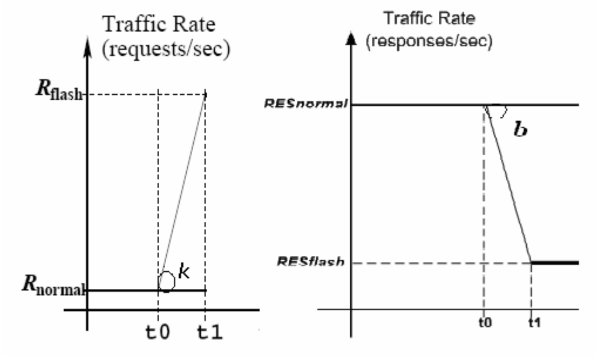

Figure 3: The k and b angles on the request ramp-up and response ramp-down phases.

As mentioned previously in the explanation of equations (3) and (4) the state for a particular observation point is depended upon the change in the value of the parameters *m* and λ. We define *m* and λ for request ramp-up and response ramp-down equal with the tangents *k* and *b* respectively.

Therefore, for a single request rate observational point *(td, Rd)* and a single response rate observational point *(td, RESd)* we have:

$$m = \tan(k) = (Rd - Rnormal)/td - t0 \qquad \textbf{(5)}$$
$$0 \le k < 90$$

$$\lambda = \tan(b) = (RESd - RESnormal)/td - t0 \qquad \textbf{(6)}$$
$$-90 < b < 0$$

From the above equations it is important to notice that while the request ramp-up phase is taking place the value for the angle b is always negative. This is a vital observation that is used later on.

In order to accurately predict the FC, it is vital to propose certain thresholds in the values for both tangents. Based upon the trace statistics from [11, 13], we propose an initial *k* angle threshold as $0 \le k \le 30$ and *td =2sec.* In [11, 13] by having k in that range there was no threat for a sharp and continuous decrease in the response rate and having *td = 2sec* there was a composition of a satisfactory single observation point for prediction on both requests and responses. In parallel, [11, 13] traces were helpful to also determine the "safety" threshold for our *b* angle as $-30 \le b < 0$. We call it as a safety threshold because in a large range of time, responses sporadically present a negative increase but they are not related with the overall FC phenomenon. Of course, these parameters can vary and are depended upon other factors such as the number of clients where that subsequently affects the estimation of the *Rnormal* constant.

Our prediction mechanism is decomposed in to three steps. First we define a "low severity" step where our mechanism starts being suspicious if there is a possibility that a FC might occur, second we denote a "medium severity" step where conditions mainly based on the response rate state accompanied in parallel by conditions satisfied on the request rate state are checked in smaller time intervals and finally the third and most critical step that concludes the existence of a FC and likewise we name as "critical severity" step. After that final step our mechanism assures us that a FC is detected and the network is alarmed. We show the properties for each step in mathematical terms and using next to them we explain the conclusion for each step in a scenario that a FC occurs.

Low Severity Conditions

$$\left.\begin{array}{l}\tan(k)=m\\ 30<a\tan(m)<90\end{array}\right\}\quad\text{conclusion: “medium severity”}$$

Medium Severity Conditions

$$\left.\begin{array}{l}45<a\tan(m)<90\\ \tan(b)=\lambda\\ -30\leq a\tan(\lambda)<0\end{array}\right\}\quad\text{conclusion: “critical severity”}$$

Critical Severity Conditions

$$\left.\begin{array}{l}tf=(t0-td)/4\\ Ct=5\\ 50\leq a\tan(m)<90\\ -45\leq a\tan(\lambda)<0\end{array}\right\}\quad\text{conclusion: “trigger FC alarm!”}$$

As these three steps show, they represent relational conditions between the inverse tangents of the *m* and λ parameters that determine the change factor of the states described in equations (2) and (3). We introduce in the critical severity step the parameter *Ct* that represents the threshold for the counts of measurements taken while the interval of the observational time range *[t0-td]* gets smaller with the value *tf*. This happens in order to facilitate a more detailed measurement and determine if the shape on each state increases or decreases dramatically. Within Ct which is again a tuneable threshold there is a conclusion if the state of each model still holds its linear properties where this satisfies our theoretical assumptions for the ramp-up and ramp-down phases on Ψt and Vt .

D. The overall detection architecture

The prediction methodology described in the previous section composes the logic unit for our overall detection architecture. Our overall detection architecture is considered by us as complete with the inclusion of three more units. Two out of the three units are responsible for gathering request and response rate measurements respectively. The logic unit in our architecture has a runtime dynamic interaction with these two components in order to function properly and further notify the fourth unit that triggers the actual alarm in a case of a FC. This architecture is composed in a pluggable fashion providing the flexibility for the addition of new components in the case of specialized measurements that anyone would like to take and furthermore our logic unit may be extended with more sub-components in order to improve its prediction accuracy. For instance there can be a sub-component in the logic unit that in collaboration with the measurement units adapts the thresholds settled for FC detection.

Following is the abstract architecture of our overall detection mechanism stating its most basic processes again in a step mode fashion.

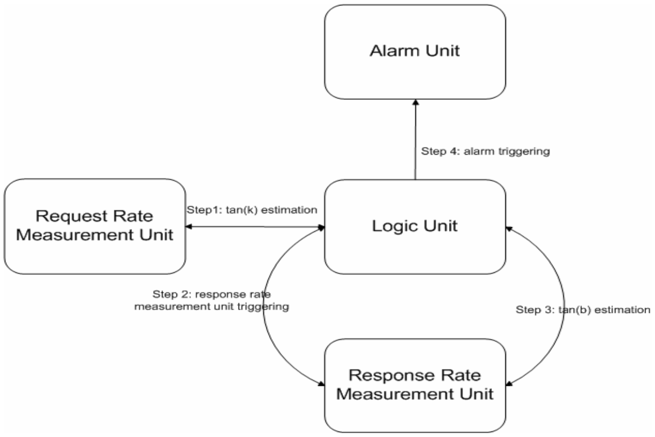
Alarm Unit
Step 4: alarm triggering
Request Rate Measurement Unit
Step1: tan(k) estimation
Logic Unit
Step 2: response rate measurement unit triggering
Step 3: tan(b) estimation
Response Rate Measurement Unit

Figure 4: The Detection Architecture

Due to the point that this document deals mainly with the prediction and the overall detection phase of the specific event, we assume that after the Alarm Unit is triggered there would be a call to another Unit in order to confront the event and mitigate it (i.e. perform load balancing).

IV. FUTURE WORK

Our proposed architecture will be initially evaluated under a realistic simulated environment which will enable the evaluation and fine-tuning of thresholds, the refinement of mathematical assumptions in the prediction methodology, and the refinement of the overall working models. Of course, simulating an unpredictable event is always a great risk. It is therefore a necessity that after such an experiment to compare our results with real traces coming from real ISPs or any kind of Autonomous Systems.

V. Conclusions

In this paper we presented an architecture for FC detection. The architecture was created by having as a basis the scenario of a single ISP. Throughout the document, we state our assumptions regarding the behavior of the responses sent and the requests received on the FC hot-spot. We have achieved to merge using a modeling approach the relationship between these two completely opposite behaviors following a linear state-based methodology. Our prediction methodology contributes in the area of identifying the changing factor of a FC which we show that is the actual slope in the both linear state models. Estimation and comparison of these slopes on the edge(s) of an ISP will enable a knowledge related to the evolutionary traffic behavior and further detection of a possible FC.

VI. REFERENCES

[1] Niven L., Flash Crowd. "***In The Flight of the Horse.***" Ballantine Books, 1971

[2] Jung J., Krishnamurthy B. and Rabinovich M., "***Flash Crowds and denial of service attacks: Characterizations and implications for CDNs and web sites.***" in WWW-02, May 2002, Hawaii, USA

[3] MSNBC website: ***http://www.msnbc.msn.com***

[4] Ari I., Hong B., Miller E., Brandt S., Long D., "***Modeling, Analysis and Simulation of Flash Crowds on the Internet***", Storage Systems research Centre, Jack Baskin School of Engineering, University of California, February 2004, Santa Cruz, CA, USA

[5] Stading T., Maniatis P., Baker M., "***Peer-to-peer caching schemes to address flash crowds.***" In proceedings of IPTPS2002, pages 203-212, March 2002, Cambridge, MA, USA

[6] Barford P., Kline J., Plonka D., Ron A., "***A Signal Analysis of Network Traffic Anomalies***" in IMW'02 , November 2002, Marseille, France.

[7] Soule A., Salamatian K., Taft N., "***Combining Filtering and Statistical Methods for Anomaly Detection***." in Internet Measurement Conference (IMC) 2005, Berkley, CA, USA.

[8] Soule A., Lakhina A., Taft N., Papagiannaki K., Salamatian K., Nucci A., Crovella M., Diot C., "***Traffic Matrices: Balancing Measurements, inference and modelling.***" In ACM SIGMETRICS (2005), ACM Press.

[9] Kalman R. E., "***A New Approach to Linear Filtering and Prediction Problems***", Instruments and Regulators Conference of the American Society of Mechanical Engineers, March-April 1959, USA.

[10]CISCO IOS NetFlow website: http://www.cisco.com/en/US/products/ps6601/products_ios_protocol_group_home.html

[11] Chen X., Heidemann J., "***Flash Crowd Mitigation via Adaptive Admission Control Based on Application-Level Observations***, ACM Transactions on Internet Technology, Vol. 5., No. 3, pages 532-569, August 2005.

[12] Ari I., Hong B., Miller E., Brandt S., Long D., "***Managing flash Crowds on the Internet***", IEEE/ACM (MASCOTS'03) October 2003, Orlando, FL, USA

[13] Chen X., Heidemann J., "***Experimental Evaluation of an Adaptive Flash Crowd Protection System***", ISI-TR- 2003-573, July 2003