Brief Announcement—Cataclysm: Handling Extreme Overloads in Internet Services *

Bhuvan Urgaonkar and Prashant Shenoy
Department of Computer Science,
University of Massachusetts, Amherst MA 01003.
bhuvan@cs.umass.edu, shenoy@cs.umass.edu

ABSTRACT

In this paper we present Cataclysm, a comprehensive approach for handling extreme overloads in hosted Internet applications. The primary contribution of our work is to develop an overload control approach that brings together admission control, dynamic provisioning of platform resources, and adaptive degradation of QoS into one integrated system. We implement a prototype Cataclysm hosting platform on a Linux cluster and demonstrate the benefits of our integrated approach using a variety of workloads.

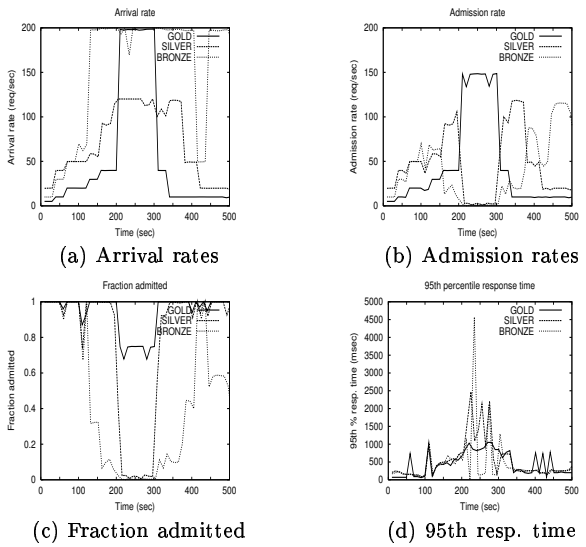
Categories and Subject Descriptors: D.4.7 [Organization and Design]: Distributed Systems

General Terms: Algorithms, Design, Experimentation.

Keywords: Internet Service, Hosting Platform, SLA, QoS, Admission Control, Dynamic Provisioning, G/G/1 Model, Linux.

1. KEY CONTRIBUTIONS

Internet applications are known to experience extreme overloads during which the workload unexpectedly increases by upto an order of magnitude in a few tens of minutes [1]. The key contribution of our work is to integrate techniques such as admission control, dynamic resource provisioning, and adaptive degradation of QoS into one integrated system for handling such overloads. We propose very low overhead admission control mechanisms that can scale to very high request rates under overloads. Our mechanisms can preferentially admit important requests during an overload and scale to handle incoming rates of up to a few tens of thousands of requests/s. Our dynamic provisioning mechanism employs a G/G/1-based queuing theoretic model of a replicable application in conjunction with online measurements to dynamically vary the number of servers allocated to each application. Further, the admission controller and provisioning mechanism cooperate with one another, and thereby enhance the ability of the platform to counter overloads.



(a) Arrival rates (b) Admission rates

(c) Fraction admitted (d) 95th resp. time

Figure 1: Working of the admission controller during an overload.

We have implemented a prototype Cataclysm hosting platform on a cluster of Linux servers. We demonstrate the effectiveness of our integrated overload control approach via an experimental evaluation. Our results show that (i) preferentially admitting requests based on importance and size can increase the utility and effective capacity of an application, (ii) our provisioning is both agile and effective at diverting platform resources to where they are needed most, thus improving platform revenue. Figure 1 shows the effectiveness of Cataclysm in providing preferential admission to more important requests for an overloaded Internet application. More results of our experimental study and the details of the design and implementation of Cataclysm can be found in [2].

2. REFERENCES

[1] The Internet Under Crisis Conditions: Learning from September 11. Committee on the Internet Under Crisis: Learning from September 11, National Research Council, 2003

[2] B. Urgaonkar and P. Shenoy. Cataclysm: Handling Extreme Overloads in Internet Services. Technical Report TR03-40, Department of Computer Science, University of Massachusetts, December 2003.

*This research was supported in part by NSF grants CCR-9984030 and EIA-0080119.

PODC'04, July 25–28, 2004, St. Johns, Newfoundland, Canada.
ACM 1-58113-802-4/04/0007.