Flash Crowds & Denial of Service Attacks

Characterization and Implications for CDNs and Web sites

Jaeyeon Jung

MIT Laboratory for Computer Science

Balachander Krishnamurthy and Michael Rabinovich

AT&T Labs-Research

Motivation

- ✔ Flash crowd is a sudden, large surge in traffic to a particular Web site
  - September 11, Ken Starr's report, Victoria's Secret webcast
- ✔ Denial of Service (DoS) attack is an explicit attempt to prevent legitimate users of a service from using that service
  - HTTP request flooding, attack to crack password-protected web pages, `Code Red` worm, TCP SYN flooding, etc.

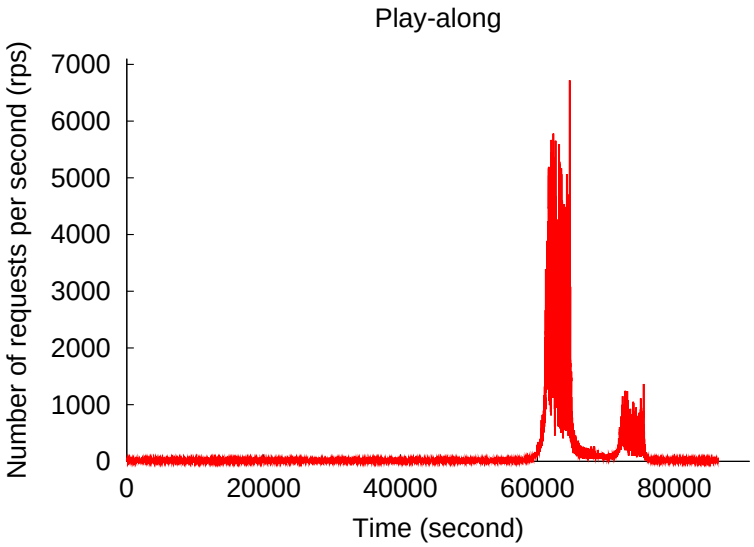
Questions

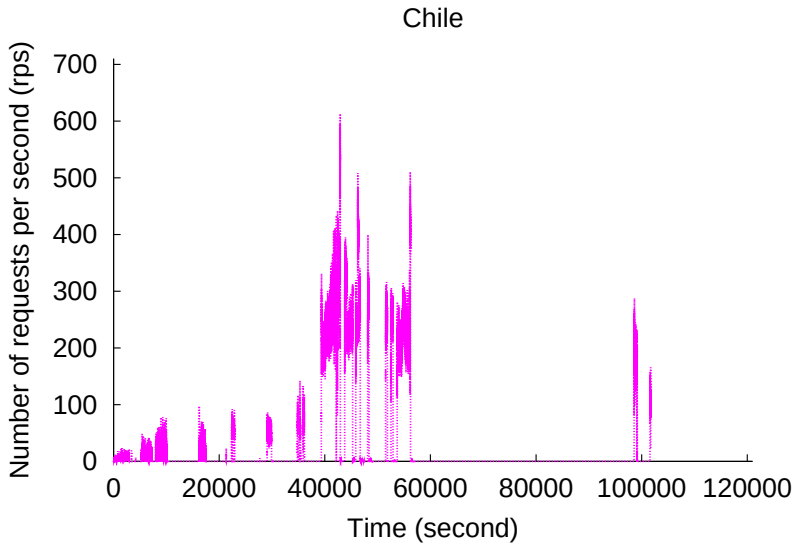
- ✔ Part I: Flash events vs. DoS attacks
  - What properties differentiate DoS attacks from flash events?
  - How can we use them to identify and separate DoS attacks from flash events?
- ✔ Part II: Flash crowds and CDNs
  - What is the locality of file reference like during flash events and its implication for CDNs?
  - How can we improve protection of Web servers from flash crowds using CDNs?

Network-Aware Clusters [KW00]

- ✔ Clustering uses a large collection of unique network prefix from BGP tables
- ✔ Classify all the IP addresses that have the same longest matched prefix into a cluster
- ✔ It helps determine topological distribution of clients in FE and DoS

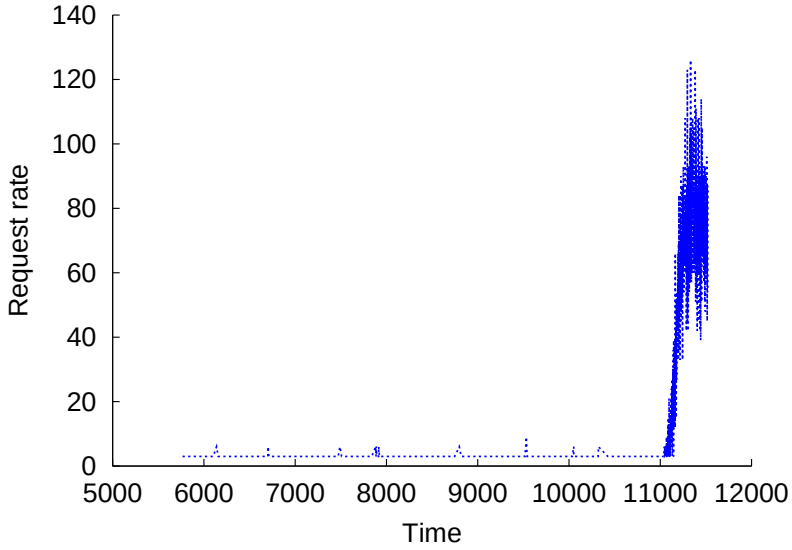
Flash Events





Trace		Requests	Documents	Clients	Clusters
Play-along	Total	13,018,385	7,084	53,745	14,100
	FE	71.0%	68.9%	63.9%	61.6%
Chile	Total	2,634,567	10,302	20,532	1,739
	FE	88.2%	90.2%	89.0%	86.6%

DoS Attacks

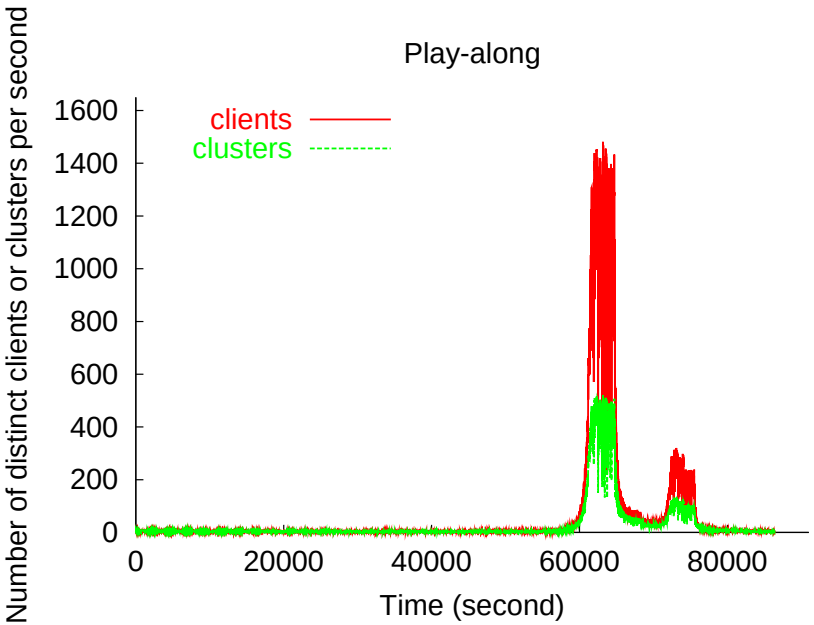


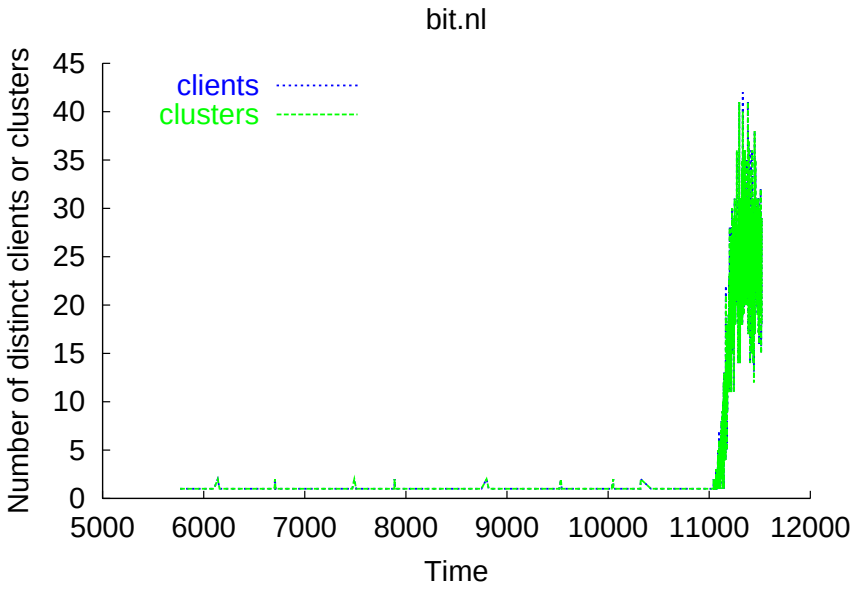
Trace	Requests	Documents	Clients	Clusters
bit.nl	35,657	1	11,092	6,155

- `Code Red` : In the earlier variant, each instance uses the same random number generator seed to create the list of IP addresses it scans [CERT].

Part I: Flash Events vs. DoS Attacks

Client Characteristics





- ✔ [FE] Clients can be effectively aggregated into clusters
- ✔ [DoS] Distribution of DoS attackers is broad

Client Characteristics - *contd.*

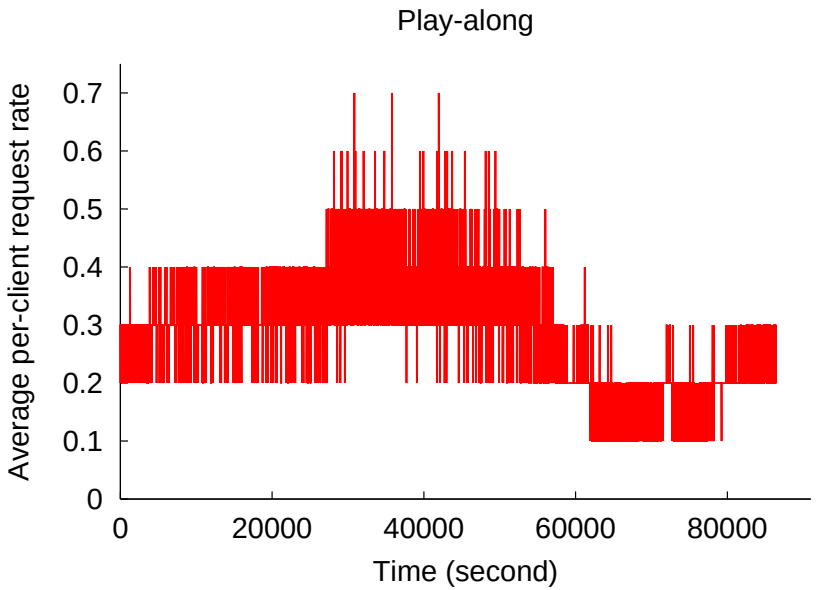
- ✔ [FE] Many *old* clusters are represented in flash events

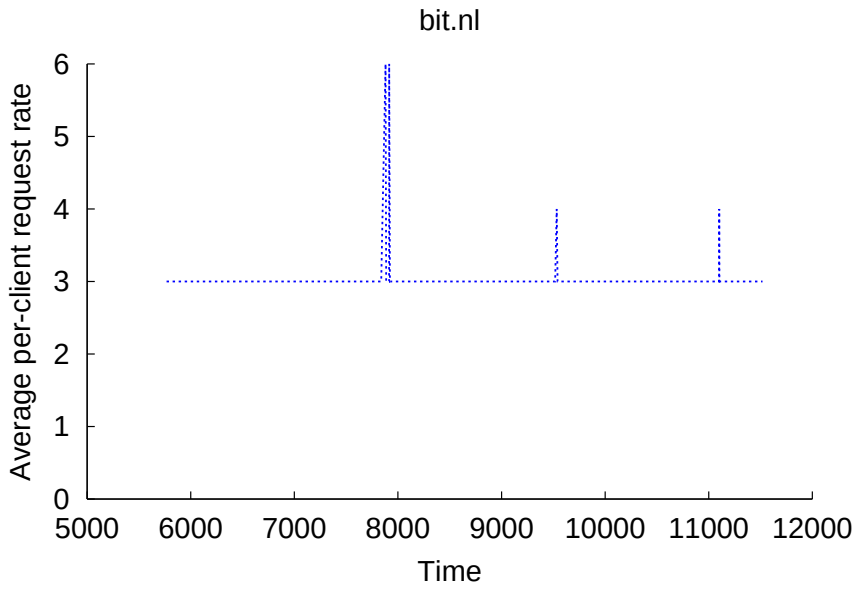
  Play-along : 42.7% and Chile: 82.9%

- ✔ [DoS] Very few previously seen clusters are involved in DoS attacks

  creighton: 0.6%, fullnote: 0%, spccctxus: 1.8%, and rellim: 14.3%

Per-client Request Rate





- ✔ [FE] There is a *decline* in per-client request rate during the flash event
- ✔ [DoS] The per-client request rate does not change during the surge in requests

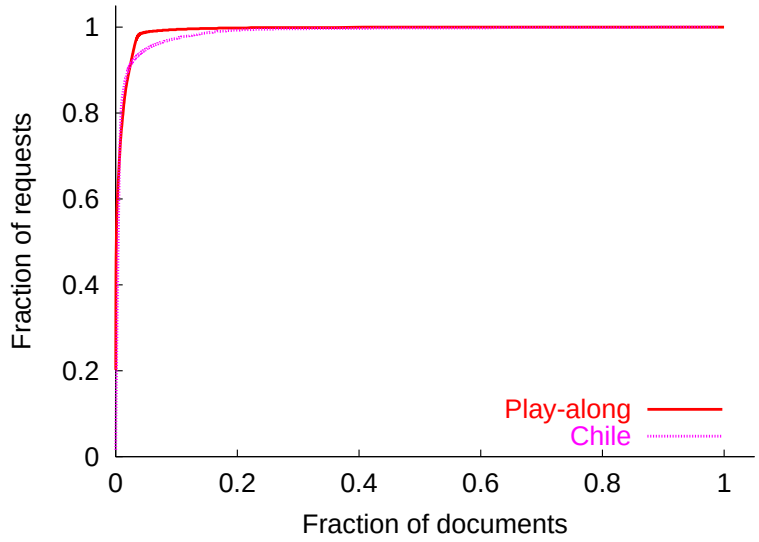
Server Strategy

- ✔ Monitor the clients that access the site and their request rate
- ✔ Periodically perform network aware clustering over the client set accumulated over the past period without flash or DoS events - *old* clusters
- ✔ When performance degrades to a threshold level, discard packets that come from clients that do not belong to *old* clusters as well as from non-proxy clients whose request rate deviates significantly from average

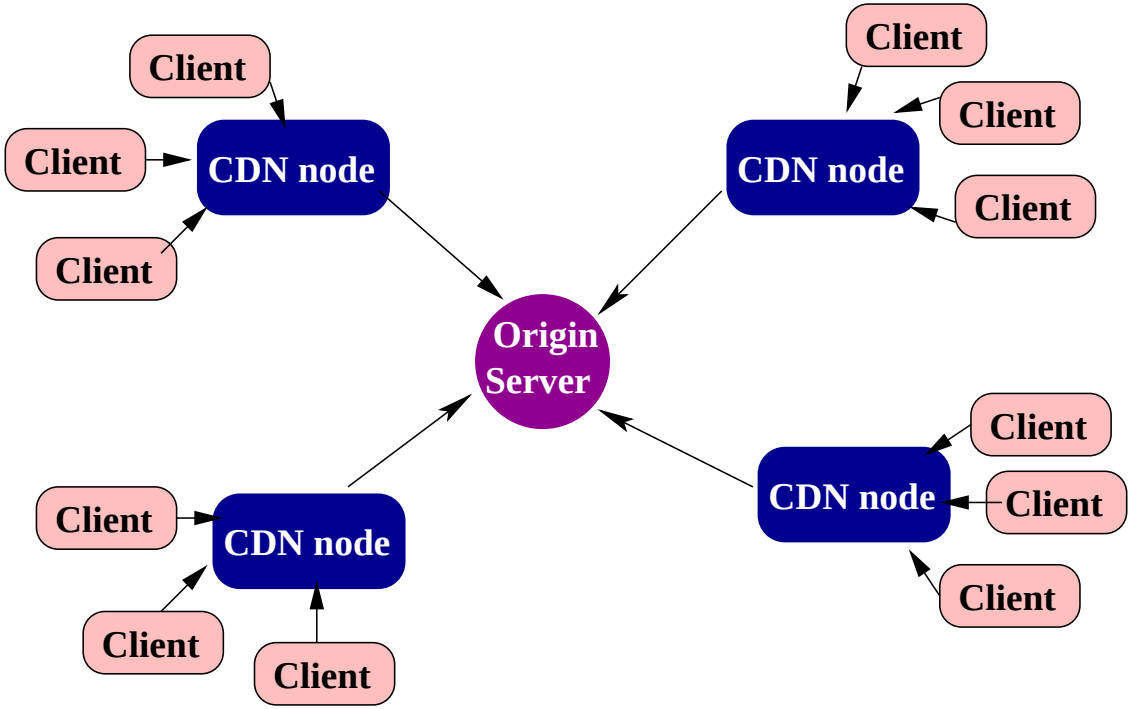
Part II: Flash Crowds and CDNs

File Reference Characteristics

- ✔ Large number of documents are accessed *only* during FEs (Play-along : 61% and Chile: 82%)
  - ➔ Many cache misses at the beginning of FEs
- ✔ 10% of popular documents account for more than 90% of requests.

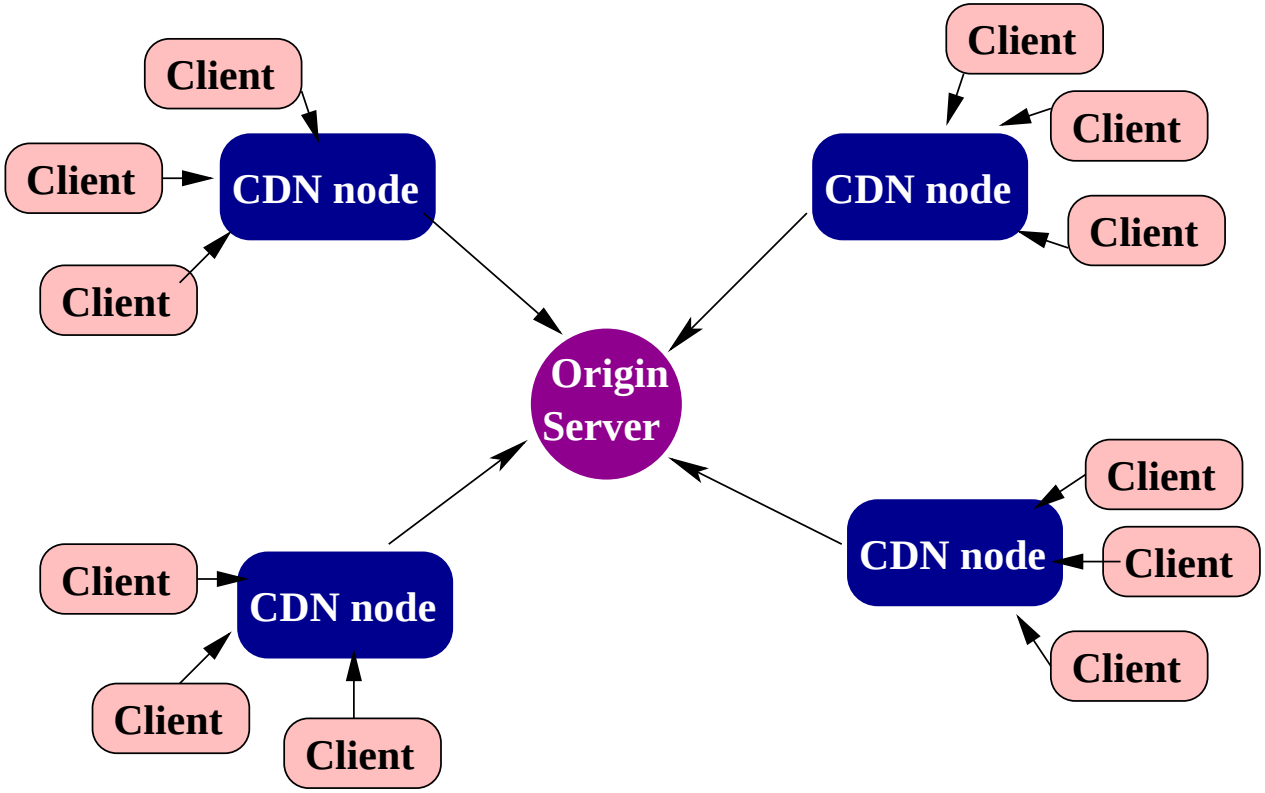


Flash Events and CDN



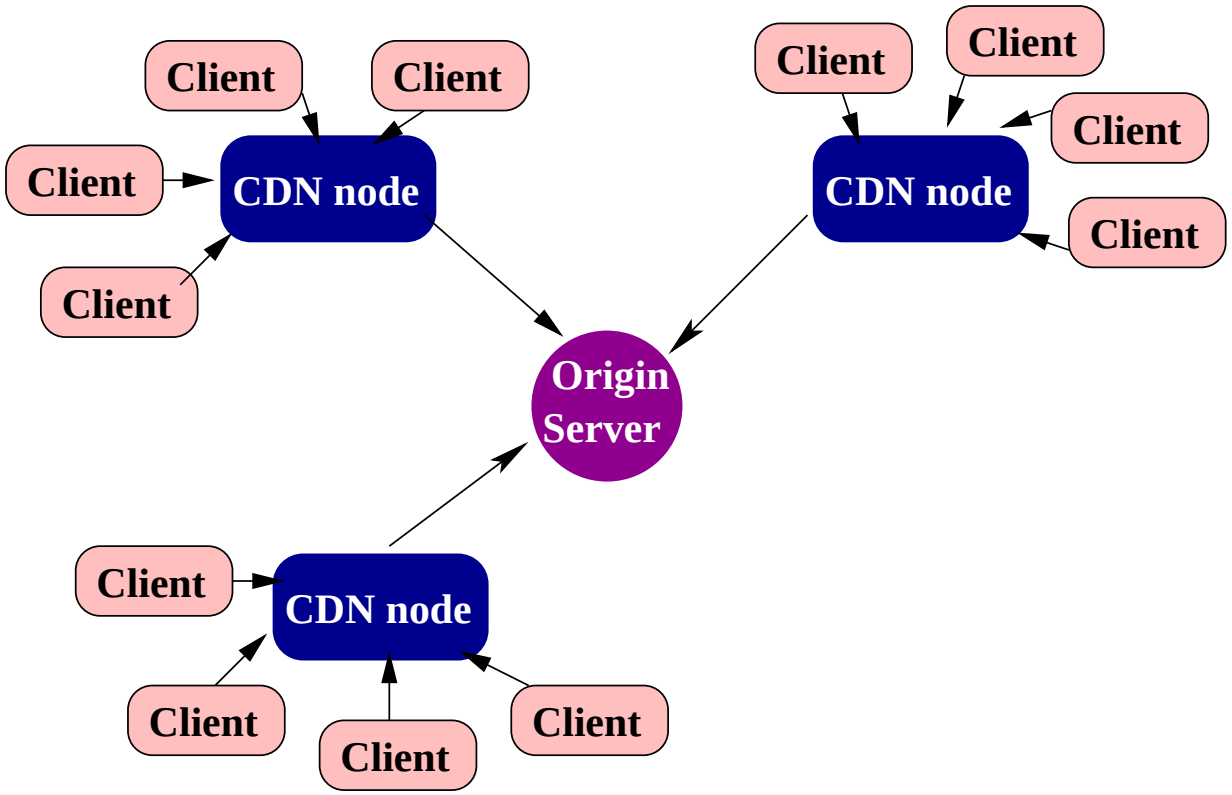
- ✔ CDN with 1,000+ cache nodes might not be able to provide an absolute protection against FE due to the peaks in the beginning of FE

Flash Events and CDN



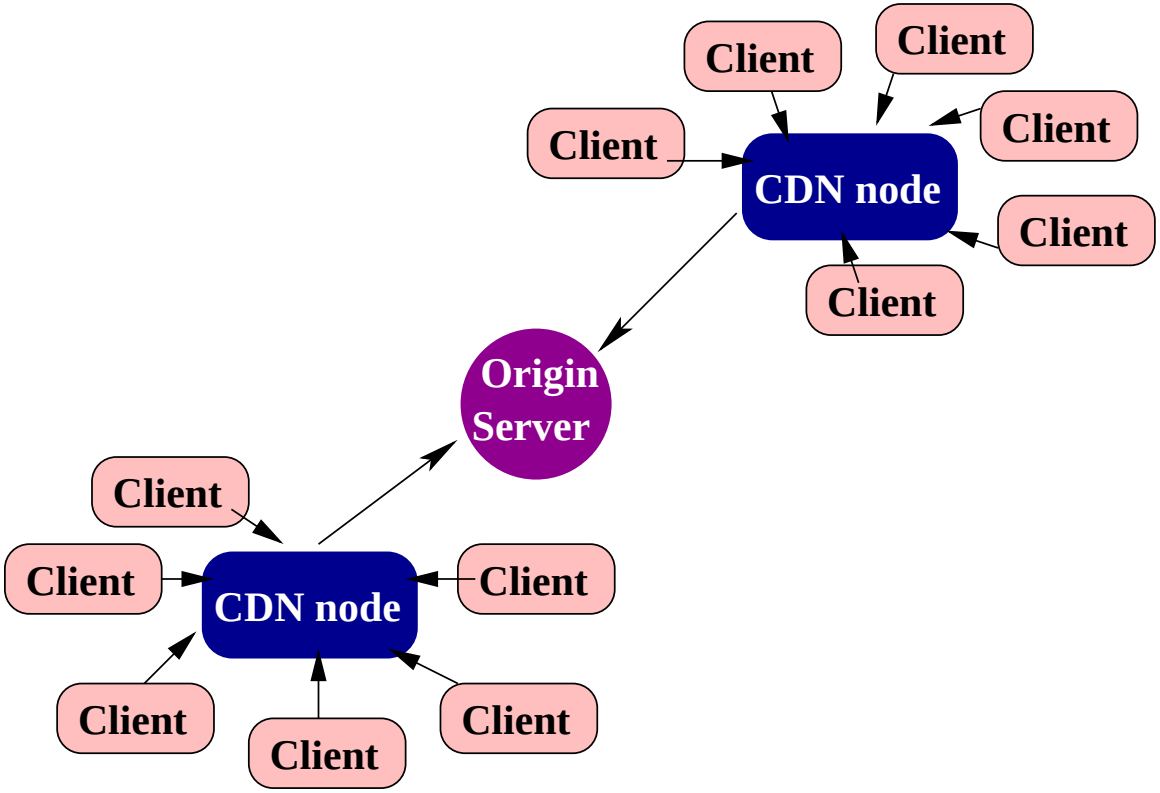
- ✔ Limiting the number of caches would increase the load on individual caches

Flash Events and CDN



- ✔ Limiting the number of caches would increase the load on individual caches

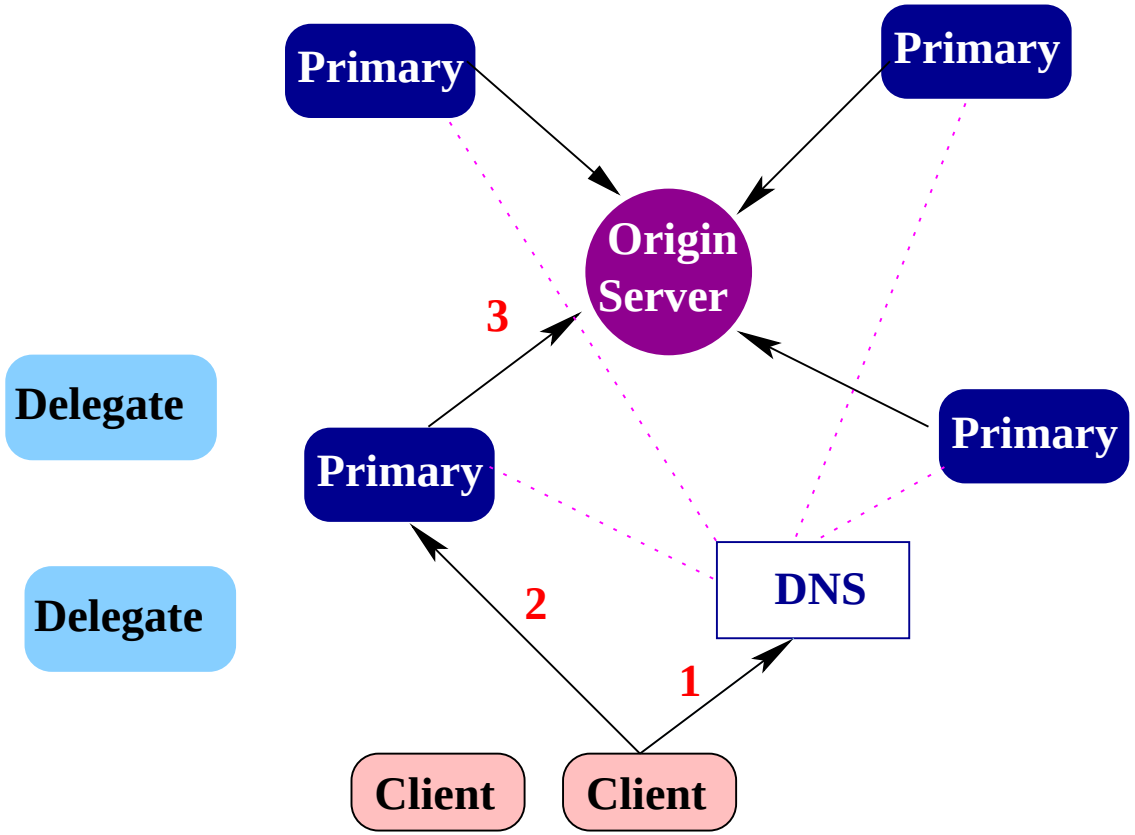
Flash Events and CDN



- ✔ Limiting the number of caches would increase the load on individual caches

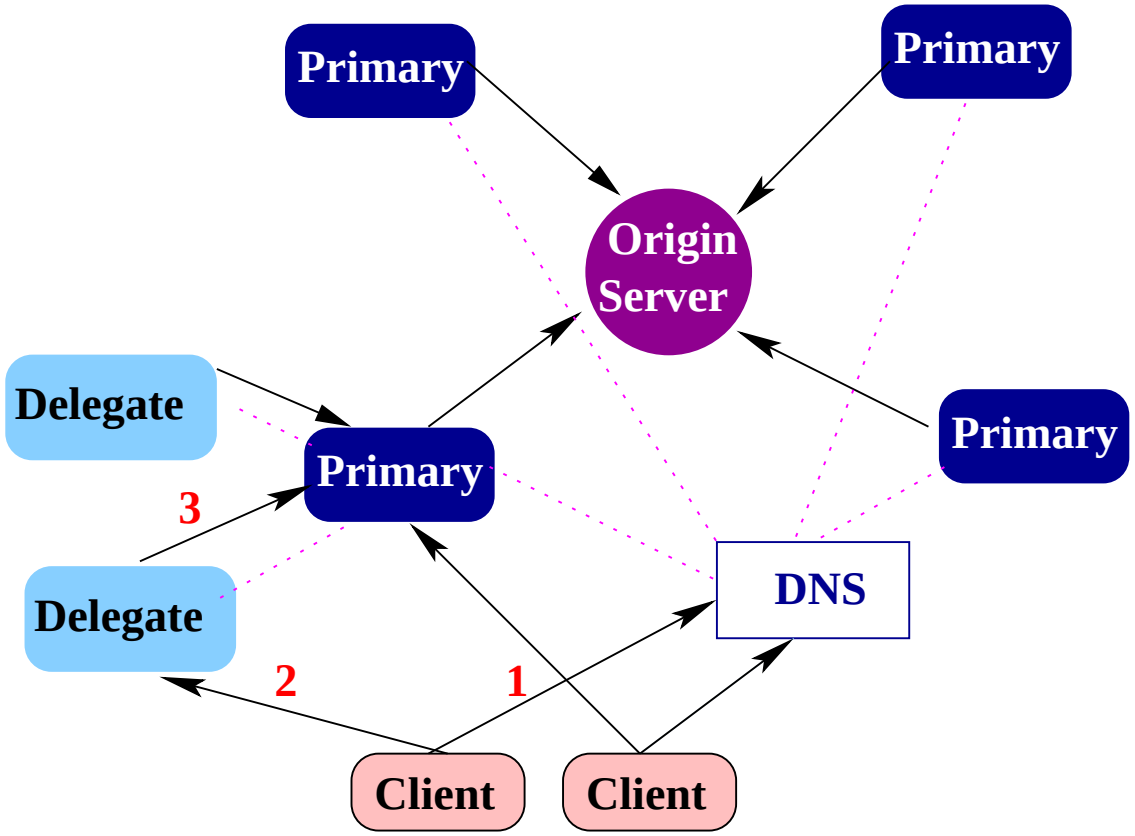
Adaptive CDN

- ✔ Lower the peak rate forwarded to an origin server while spreading load over cache nodes



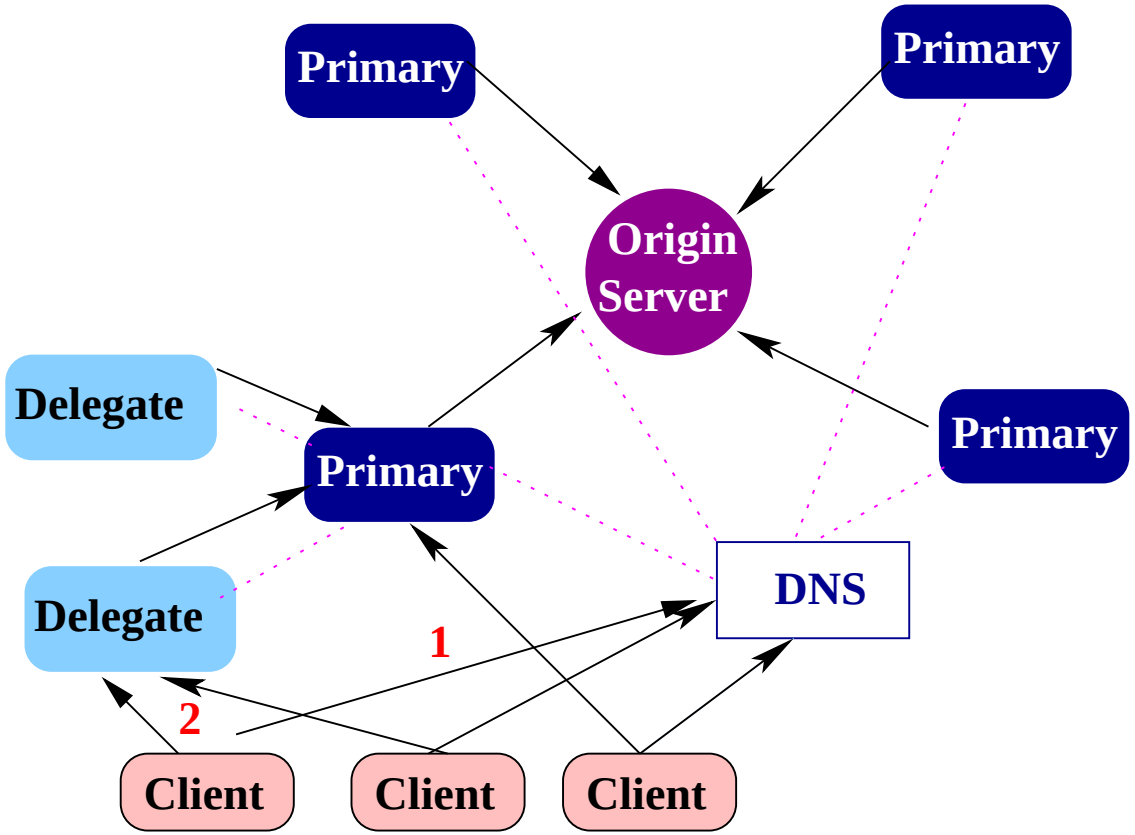
Adaptive CDN

- ✔ Lower the peak rate forwarded to the origin server while spreading load over cache nodes

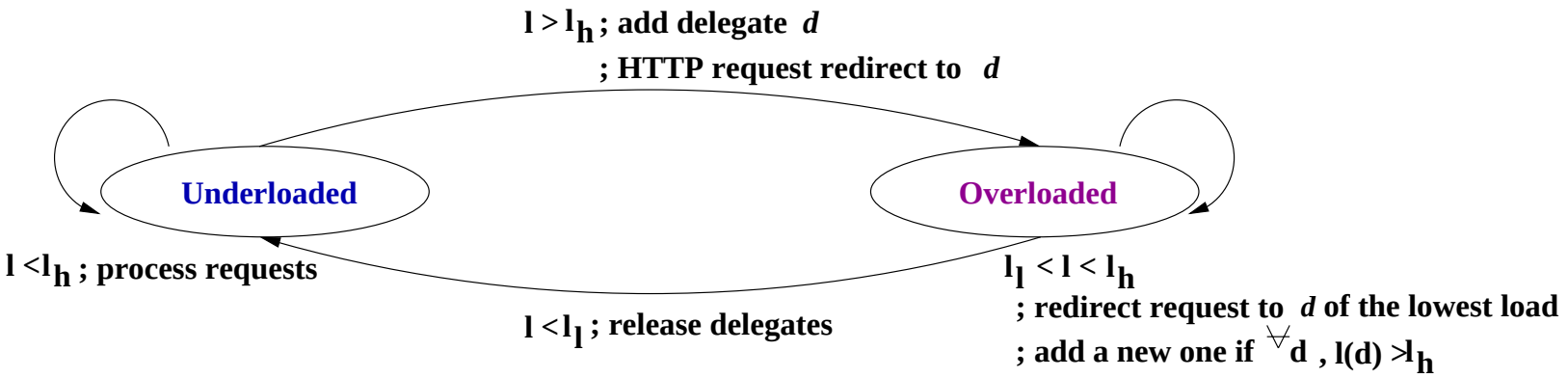


Adaptive CDN

- ✔ Lower the peak rate forwarded to the origin server while spreading load over cache nodes

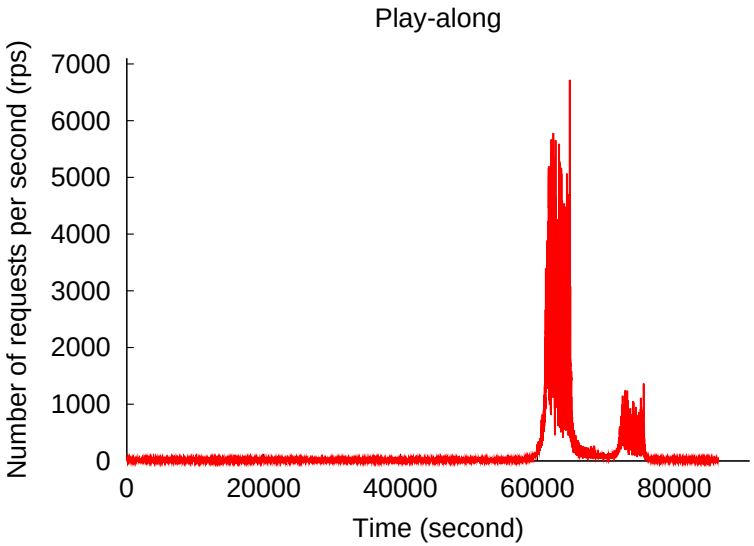


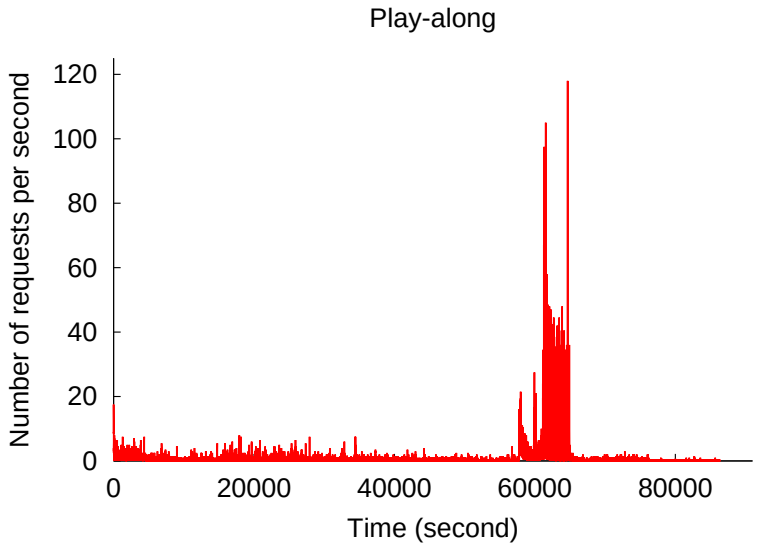
Dynamic Delegation



- ✔ Load on primary cache, *l*, is computed as requests per second averaged over two-second interval.
- ✔ Cache is assigned based on cluster and used for *ttl* period.

Simulation Results





- ✔ Peak request rates are reduced by a factor of 50 (Play-along), and 20 (Chile)
- ✔ It ensures that load on each cache remains low (50 rps) and that proximity-based cache selection is not compromised.

Conclusion

- ✔ Client clustering technique is useful for source identification and for distinguishing legitimate requests and malicious attacks.
- ✔ Per-client request rate drops and remains lower during the FEs unlike DoS attackers who generate requests independently of a server load
- ✔ Adaptive CDN is effective in terms of reduction of flows from the main server and dynamic load distribution over cache nodes.