Optimum Strategy of Cache Pollution Attacks Targeting CDN Caches

Jiaqi Liu Noriaki Kamiyama

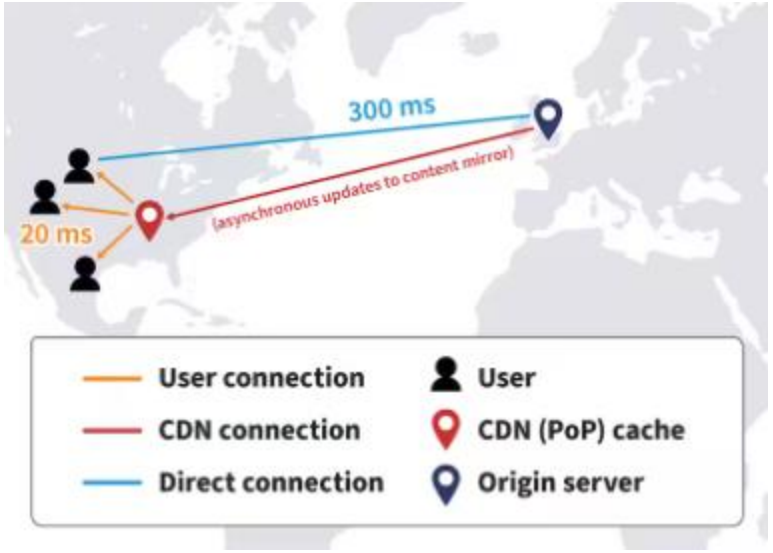
Ritsumeikan Univ

2025.03.14

Content Delivery Network

- Content Delivery Network (CDN)
  - Origin servers: Provide the original version of the content
  - Cache servers: Cache the copy of contents, and they are responsible for delivering that content to nearby users.
  - DNS servers: Respond user's request with the name of a cache server from which the content can be served faster.

- The feature of CDN
  - Serves a large portion of the Internet content
  - Provides a faster and high-performance experience
  - Reduce bandwidth costs



Purpose of research

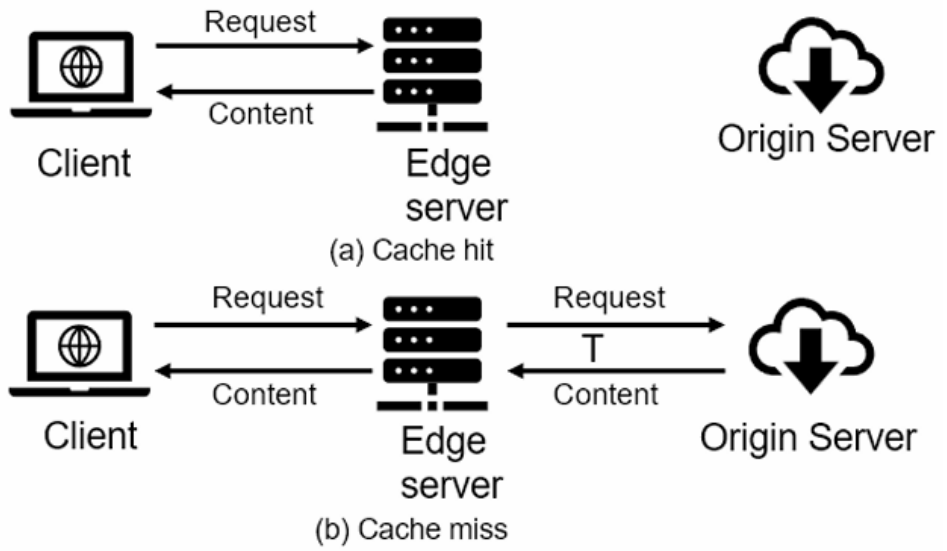
- Cache pollution attack（CPA）
  - Pollute the cache with low-popularity content to degrade the performance of the cache
- Propose the analytical model to evaluate the impact of CPA.
- Analyzes the impact of specific scenarios on CPA

Analytical Model

- M/M/1 queue
  - $W = \frac{1}{\mu - \sum_{i=1}^{M} \lambda_i}$

Parameter	Definition
W	Average response time
1/μ	Average service time
M	Number of contents
λ_i	Poisson arrival rate of request for content i
h_i	Cache hit ratio of content i

- Cache Server
  - The latency time T is spent when cache misses
  - $R(i) = h_i W + (1 - h_i)(W + W_O + T)$



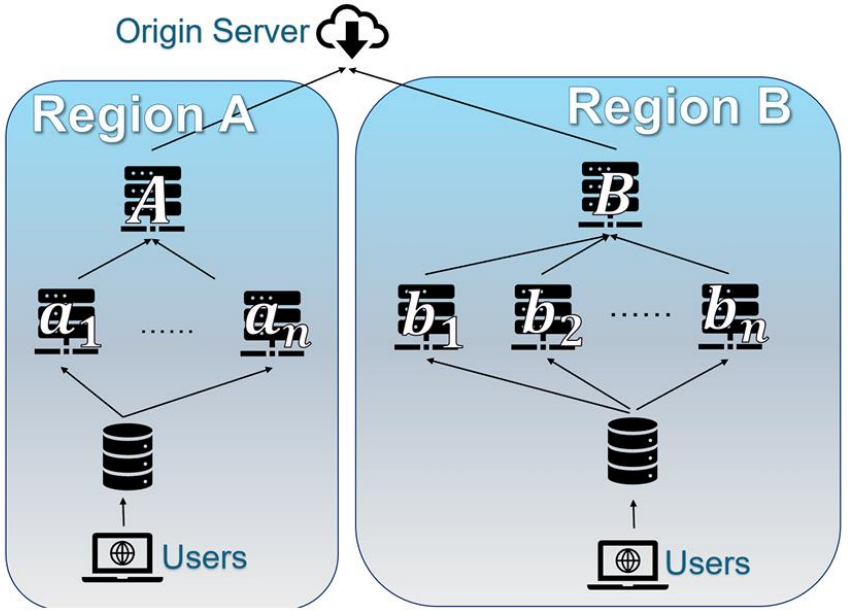
Analytical Model

- Che-Approximation
  - $h_i \approx 1 - e^{-q_i t_c}$
  - $\sum_{i=1}^{M} h_i = C$

Parameter	Definition
q_i	Request ratio of content i
C	Capacity of cache
t_c	Characteristic time

CDN Model

- The origin server provides content copies for 5 regions
- The central server has a large cache size for storing content from the origin server and sending it to the edge servers.
- The edge server is responsible for quickly providing popular content with a small cache size.
- DNS server is responsible for randomly locating user requests to edge servers
- All server adopt LRU
  - Least Recently Used



Multilayer CDN Model

- Average response time in Region A
  - $r_e^A = W_a$ -Cache hit in a_n
  - $r_c^A = W_a + W_A + T_1$ -Cache hit in A
  - $r_o^A = W_a + W_A + W_O + T_1 + T_2$ -Cache miss in A and a_n
- Average response time of content i when request arrives at region A
  - $R_i^A = h_i^{an} r_e^A + (1 - h_i^{an}) h_i^A r_c^A + (1 - h_i^{an})(1 - h_i^A) r_o^A$
- Average response time of all requests in region A
  - $R_A = \sum_{i=1}^{M} \frac{\lambda_i^A R_i^A}{\sum_{i=1}^{M} \lambda_i^A}$

Attack definition

- Cache pollution attack (CPA)
  - CPA will make many requests for low-popularity content, which will decrease the cache hit ratio of popular content and result in a longer response time for most users, especially in CDN
  - Generally, it's difficult to distinguish a request sent by CPA from legitimate requests, so we assume that the request sent by CPA cannot be detected

Attack definition

- Optimize Attack
  - To optimize attack, attacker will maximize the effect of the attack by adjusting the proportion of requests to different contents in different servers and observing the impact on the response time.
  - The attacker can only approach the optimal attack strategy by keep observing and adjusting.

Genetic Algorithm

- Genetic Algorithm (GA), a kind of evolutionary algorithms, is inspired by the process of natural selection.
  - Mutation
  - Crossover
  - Selection

- With GA, we can get the theoretically optimal attack strategy, which the attacker can approach infinitely but not exceed.

Genetic Algorithm

- Workflow of GA
  1. Randomly set s_i^n as the initialized chromosome, which 0 < s_i^n <1
  2. To achieve crossover, set any two chromosomes as a group and randomly exchange of s_i^n in groups.
  3. To achieve mutation, randomly change a s_i^n, which 0 < s_i^n <1
  4. To achieve selection, add attacker's request to the normal request and calculate the average of R as fitness. Then select the best chromosomes from the parents and children.
  5. Repeats steps 2, 3 and 4 until N generations.

- In the end, we obtained a series of s_i^n.
  - The attacker multiplies the total number of requests they can use by s_i^n to obtain the request rate sent to content i in region n.

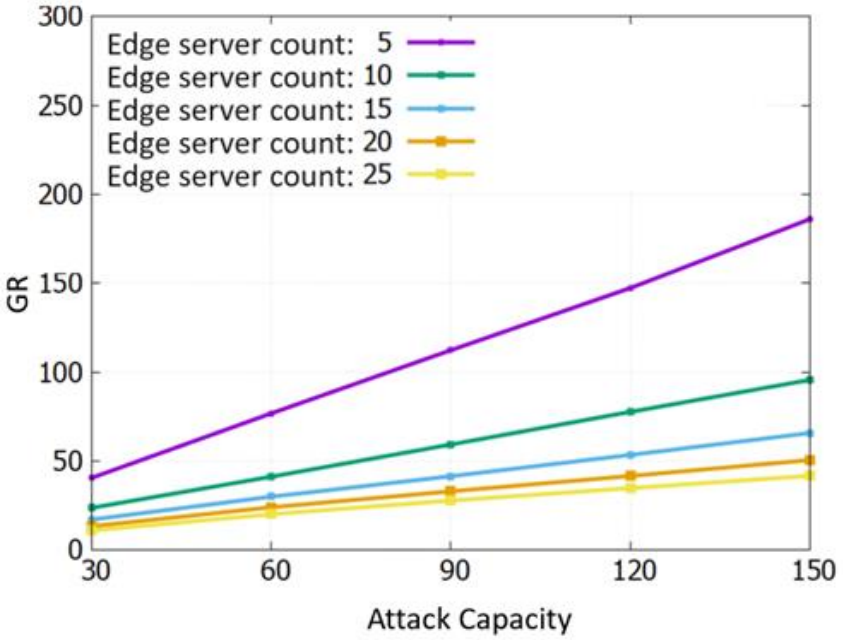
Evaluation

Parameter	Value
Number of contents provided, M	100
Edge servers' cache size	20
Central servers' cache size	50
Zipf law parameter	4
Requests rate to each edge server, λ	30 /second
Av. service time of Edge servers, 1/μ1	10 ms
Av. service time of Central servers, 1/μ2	2 ms
Av. service time of origin server, 1/μo	1 ms
Latency, T1	500ms
Latency, T2	300ms

$$GR = \frac{R_A - R_A^n}{R_A^n}$$

Evaluation: Edge Server Count

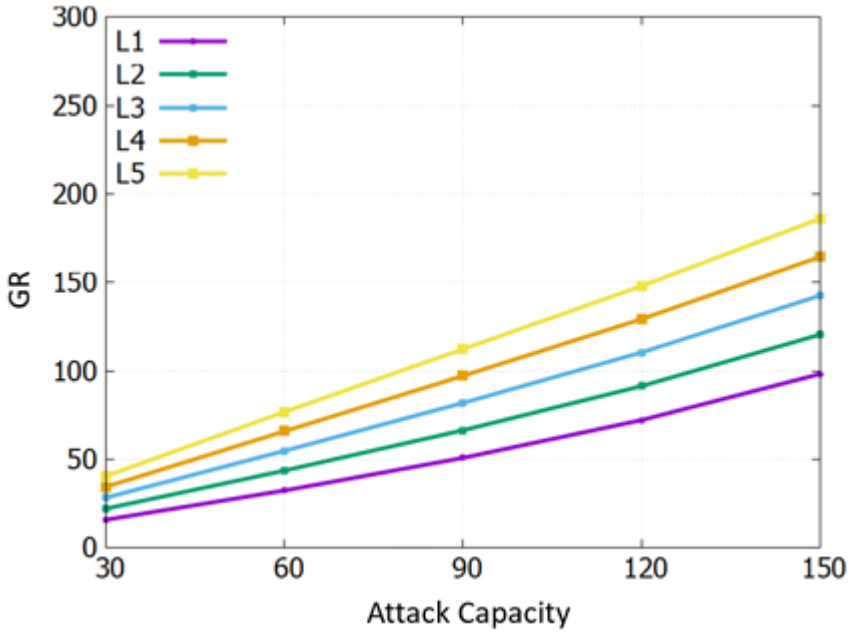
- With the same attack capacity, region with less edge server count are more vulnerable to attacks
- As the attack capacity increases, the growth rate of GR in region with lower number of edge servers increases



Evaluation: Latency

- When the attack capacity is small, the impact of latency on CDN is minimal because CDN's caching mechanism can effectively reduce the impact of latency on response time

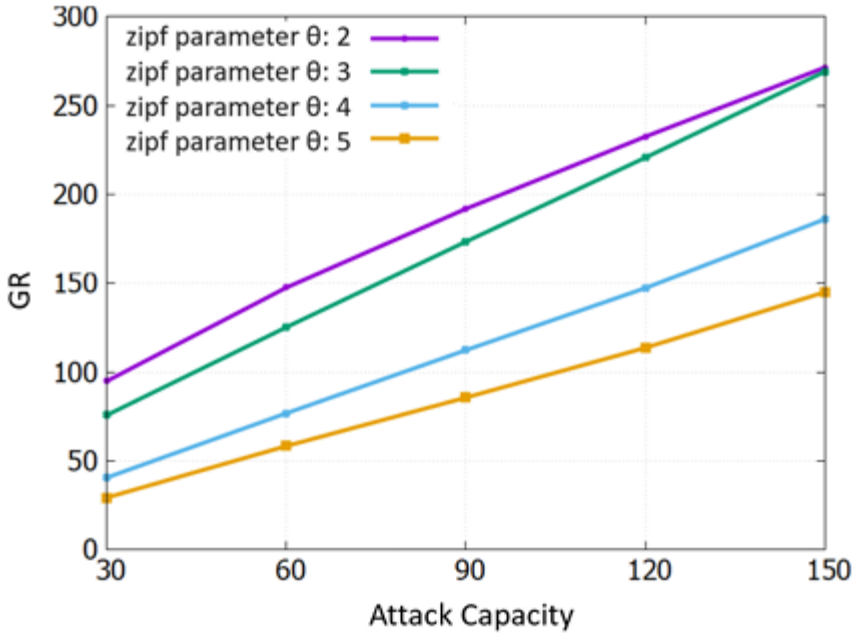
Parameter	T_1	T_2
L1	60 ms	100 ms
L2	120 ms	200 ms
L3	180 ms	300 ms
L4	240 ms	400 ms
L5	300 ms	500 ms



Evaluation: Zipf's Law Parameter

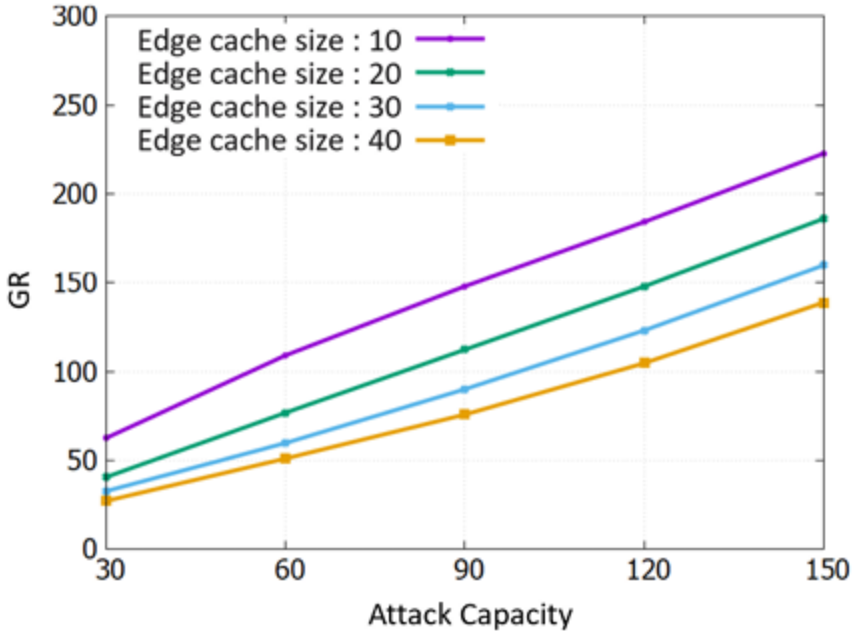
- The higher the Zipf's Law parameter, the more concentrated the popularity of the content becomes.

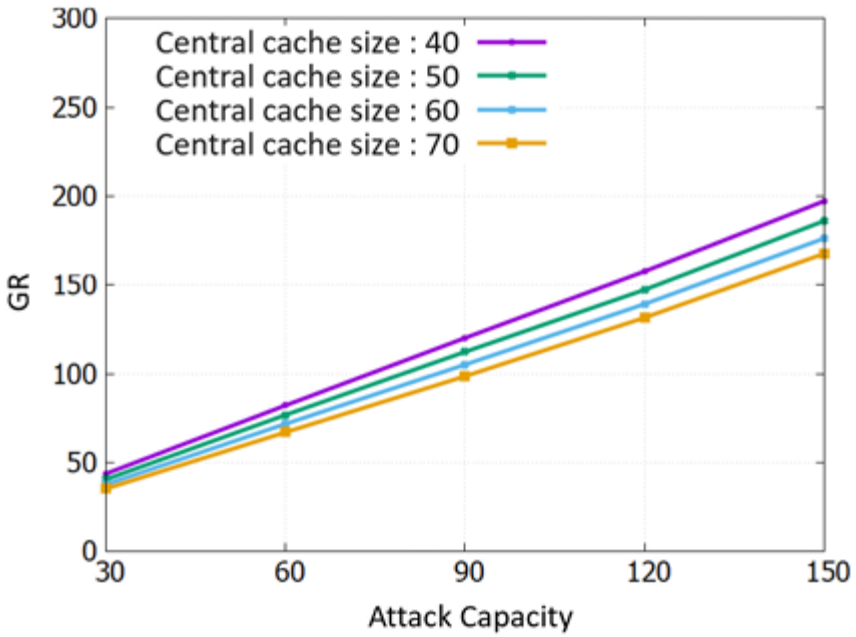
- Content with more dispersed popularity is more susceptible to cache attacks



Evaluation: Cache Size

- Edge cache size is more sensitive to CPA
- Central cache size has a small impact on CPA





Conclusion

- We used the M/M/1 queue model to derive the response time for server in CDN
- We build a multi-region CDN model according to the actual CDN, and compared the GR under different attacks
- We investigated different scenario and revealed the impact of different parameter.