Midgress-aware traffic provisioning for content delivery

Aditya Sundararajan, Mangesh Kasbekar, Ramesh K. Sitaraman, Samta Shukla
CDNs serve more than 50% of content
Performance and cost metrics

End-user latency

Origin offload ratio

Bandwidth cost

Cache hit rate
100s of content providers

100s of 1000s of servers

Millions of users
Past work has focused on cache management.
How can we assign traffic classes to reduce midgress?
Traffic provisioning to reduce midgress

100s of traffic assignment scenarios!
Traffic provisioning to minimize midgress
Eviction age equality
Footprint descriptors*

**Spatial locality**: How many unique bytes are requested between successive requests of an object?

**Temporal locality**: How often is an object requested?

---

* Footprint descriptors: Theory and practice of cache provisioning in a global CDN, A. Sundarrajlan et al. in ACM CoNEXT 2017
Caching properties from FDs

- Hit rate = $f$ (size)
- Hit rate = $f$ (eviction age)
- Cache size = $f$ (eviction age)
- Joint probability distribution $P(s,t)$

Footprint descriptors: Theory and practice of cache provisioning in a global CDN, A. Sundararajan et al. in ACM CoNEXT 2017
Traffic mixing using FD calculus

The addition operation is the convolution of joint pdfs which can be efficiently computed using FFT

* Footprint descriptors: Theory and practice of cache provisioning in a global CDN, A. Sundararajan et al. in ACM CoNEXT 2017
Traffic provisioning to minimize midgress

Min. miss traffic to origin

FD of traffic classes

FD calculus to optimize traffic class assignment

Origin

CDN

Users
Traffic provisioning as an optimization problem

\[ \text{Min.} \sum_{ij} x_{ij} \lambda_j m_j(c_{ij}) \]

Total miss traffic from cluster

\( T \) traffic classes
\( \lambda_1, \lambda_2, \ldots, \lambda_T \)

Cache size, \( C \)
Traffic capacity, \( B \)

\( N \) servers

MILP – NP Hard!!

Estimate miss rate of traffic mix using \textit{FD calculus}
FD-based local search is faster than MILP

1. Randomly assign traffic classes

Traffic classes

Servers

Predict midgress of traffic mix using FD calculus
FD-based local search is faster than MILP

2. Reassign traffic classes using local search such that midgress is minimized

Predict midgress of traffic mix using FD calculus
Metro-level traffic provisioning

Traffic classes

Cluster 1

Servers

Cluster N

Servers

Midgress of metro area
## Trace characteristics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of traffic classes</strong></td>
<td>25</td>
</tr>
<tr>
<td><strong>Length of trace</strong></td>
<td>16 days</td>
</tr>
<tr>
<td><strong>Traffic types</strong></td>
<td>Web, media, download</td>
</tr>
</tbody>
</table>
Metro-level midgress reduced by 20%
Traffic provisioning in partitioned caches

![Graph showing cache miss rate (%) vs cache size (TB). The graph compares OPT, baseline fit, OPT-part, and baseline fit-part. The x-axis represents cache size in TB, ranging from 0 to 600, and the y-axis represents cache miss rate in percentage, ranging from 0 to 60].
Conclusions

Midgress-aware traffic provisioning reduced midgress by almost 20% in metro area

Midgress-aware heuristic performs within 1.1% of OPT but is much faster

Midgress-aware traffic provisioning can be extended to work with additional constraints such as minimum redundancy and maximum midgress, any cache management algorithm, and with partitioned caches
Thank you!

Email: asundar@cs.umass.edu