Semi-edge: From edge caching to hierarchical caching in network fog

This item was submitted to Loughborough University's Institutional Repository by the/an author.


Additional Information:

- This is the author’s version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in Proceedings of the 1st International Workshop on Edge Systems, Analytics and Networking (EdgeSys'18), https://doi.org/10.1145/3213344.3213352

Metadata Record: https://dspace.lboro.ac.uk/2134/33644

Version: Accepted for publication

Publisher: © Association for Computing Machinery (ACM)

Please cite the published version.
ABSTRACT

In recent content delivery mechanisms, popular contents tend to be placed closer to the users for better delivery performance and lower network resource occupation. Caching mechanisms in Content Delivery Networks (CDN), Mobile Edge Clouds (MECs) and fog computing have implemented edge caching paradigm for different application scenarios. However, state-of-the-art caching mechanisms in literature are mostly bounded by application scenarios. With the rapid development of heterogeneous networks, the lack of uniform caching management has become an issue. Therefore, a novel caching mechanism, Semi-Edge caching (SE), is proposed in this paper. SE caching mechanism is based on in-network caching technique and it could be generically applied into various types of network fog. Furthermore, two content allocation strategies, SE-U (unicast) and SE-B (broadcast), are proposed within SE mechanism. The performance of SE-U and SE-B are evaluated in three typical topologies with various scenario contexts. Compared to edge caching, SE can reduce latency by 7% and increase cache hit ratio by 45%.

CCS CONCEPTS
• Information systems → Hierarchical storage management;
• Networks → Network simulations; In-network processing.

KEYWORDS
Edge Caching, In-network Caching, Information Centric Networking, Content Delivery

1 INTRODUCTION

Nowadays, content delivery has taken the place of host-to-host communication as the main application of the Internet. Cisco’s forecast [1] shows that more than 70% traffic will be carried by content delivery networks in 2021. Moreover, the percentage of IP traffic generated by smartphones will grow to 33% while this percentage of PCs will drop to 25%. Some caching-based applications (e.g. Web Caching, CDN, P2P) have been developed to enhance the performance of content delivery services by placing content geographically closer to the users. In another point of view, the research about future networks architectures with higher mobility, reliability and scalability is motivated by the explosive development of Internet of Things (IoT) and 5G communications. Therefore, caching mechanisms in future network has become a promising direction in both academia and industry.

Cloud computing is a future network architecture which decoupling service providers from geographical locations. Cloud caching has evolved client-server content delivery into pay-as-you-go service model through cloud storage implementation. Thus, the scalability and flexibility of content delivery are improved because services requested by mobile users could be satisfied by not only original providers but also resource pools. In recent years, the position of resource pools tends to be moved from network core to the edge. Consequently, advanced paradigms like fog computing or cloudlet are sequently proposed in academia [6].

More recently, Information Centric Networking (ICN) has been rapidly developed as a novel future network infrastructure [25]. In ICN, delivered objects are named independently of location and cached within the network nodes [2]. Compared with Internet, ICN aims to implement content delivery services through a Publish/Subscribe model. Within ICN paradigm, in-network caching is one of the key features and core techniques [8].

However, to the best knowledge of the authors, most of existing ICN caching strategies are based on global network caching rather than caching only in network edge [12]. Conventional edge caching strategies are specifically bounded to either content types [26] (e.g. video, file, web page) or network types(e.g. IoT [3], WiFi, macrocell architecture). Moreover, existing literature has not sufficiently considered the effect of fog topology on caching performance. To address these issues, a novel Semi-Edge (SE) caching mechanism is proposed for network fog. SE caching mechanism can be applied not only into edge nodes but also the nodes around the edge nodes. These nodes neighbouring the edge nodes are named “semi-edge nodes”. Under SE mechanism, each user has a caching system consisted of edge and semi-edge nodes. In this paper, SE-U and SE-B are designed and evaluated as novel hierarchical caching strategies. Compared with literature, SE caching mechanism is more generic in different context scenarios because its parameters (e.g. storage capacity) are adjustable through adding or removing caching nodes.
Moreover, several novel parameters are defined to describe the topological features of network fog. The topological effect on caching performance are also evaluated through experimental results.

The remaining of paper is organised as follows: Section 2 reviews the literature on edge caching mechanisms for content delivery. Section 3 describes the proposed Semi-Edge caching mechanism (SE), including the setting up of caching system (SE), its corresponding hierarchical request routing & content forwarding strategy (in Subsection 3.1) and SE-U & SE-B strategies (in Subsection 3.2). Section 4 defines several topological features to describe fog topology. Section 5 includes the scenario configuration of simulation (in Subsection 5.1) and the comparison results of SE-U and SE-B against edge caching strategy (in Subsection 5.2). In addition, further performance evaluation about the topology and context effects are presented in Section 6.1 and 6.2 respectively. Finally, the work is summarised in Section 7 along with the future work.

2 RELATED WORK

Caching near the users is a common trend of content delivery services in both industry implementation and academia research. In industry, the caching solutions for specific content are mostly implemented in application layer, such as P2P for files [9], Web Caching for web page [4], CDN for video [23], etc. At present, not only large content providers like Google, Facebook, Netflix have already implemented their own CDN services, relatively small-scale content providers could also rent CDN services from third-party CDN companies like Akamai, Velocix, Fastly and Limelight [23]. With the widely implementation of CDN, more traffic could be delivered within a domain near the users. Although caching applications can temporarily improve user experience by alleviating content delivery burden, the complicated upper layer protocols aggravate the narrow waist of Internet.

Other than application layer solutions, a network layer solution for content delivery, Information Centric Network (ICN), was initially proposed within project TRIAD [11] in early 2000’s. Then various projects have implemented ICN in distinctly different architectures. For example, DONA [15] project by UC Berkeley was the first comprehensive design with flat naming and in-network caching contributions. Then projects like SAIL, COMET considered the feasibility of ICN implementation in current Internet architecture by introducing convergence layer and content mediation plane respectively [25]. More recently, CCN/NDN [27] with hierarchical naming scheme and PURSUIT/POINT [22] with flat naming scheme are the most active ICN branches in the research scope. Although the architectures are distinctly different between ICN branches, in-network caching is one of their common features. Similar to upper layer solutions, ICN can achieve better content delivery performance through distributed and collaborative caching mechanisms as well.

In the in-network caching literature, some existing works are related to caching deployment techniques, such as caching hop distribution [24], cache capacity allocation [18], and caching scheduling [13]. Moreover, content placement strategy and content replacement policy are also applied when particular content is supposed to be cached. The caching hops for content caching are selected according to content placement strategy. If contents copies are only cached along the delivery path from content providers to request senders [7], the content placement strategy is on-path caching. Most of content placement strategies in ICN are on-path caching, e.g. Leave Copy Everywhere (LCE), Leave Copy Down (LCD), etc [12]. Moreover, there are several advanced methodologies used for content allocation design recently, such as probability [17] and graph theory [5]. SE mechanism proposed in this paper is initially based on ICN in-network caching technique. But it could also be applied in to application layer scenarios like MEC, CDN, etc.

Regarding to edge caching, there is a lack of application compatibility in conventional literature. For example, FemtoCaching [10] is proposed for wireless 5G network without ICN features. Similarly, Tran at el. study the video-specific caching in mobile edge-computing scenario [21]. Mohan et al. present a grouping-based caching solution for IoT scenario in the cloud of edge [16] while ICN-based work by Amadeo et al. focuses on smart home services through a three layered architecture [3]. Despite of the heterogenous network types, the edge caching mechanisms for different traffic workload are also distinct because of their specific requirements. For example, although both Zhang et al [28] and Yu et al [26] are both for video delivery purpose, Zhang’s is for Video-on-Demand (VoD) workload [28] and aims to reduce the average numbers transfer hops. However, Yu’s work is for video streaming [26] aiming to avoid network congestion. Therefore, it is necessary to propose generic SE caching which can be applied into various content delivery scenarios in different network types.

3 SEMI-EDGE CACHING MECHANISM

3.1 Caching System

Semi-Edge (SE) is proposed as a generic caching mechanism in network fog. Under SE, each user is assigned a hierarchical cooperative system consisted of caching nodes in network fog. Compared with edge caching, the caching nodes in SE caching system is categorised into two levels, edge node (level 1) and semi-edge nodes (level 2) as shown in Fig.1.

In the realistic topologies of network edge, each end equipment could only connect to one edge node at one time. Hence, the amount of edge nodes in level 1 is assumed to be 1. Then the caching nodes which are one-hop distance away from edge nodes are initially configured as the level 2 participants. Every time a new receiver joins in the network, its caching system is initialised at once.

SE can be applied into most of application scenarios like IoT/mobile networks generally regardless context restrictions. Compared with universal network caching and edge caching, the scale and cache capacity of caching network in SE is relatively flexible because of the feasibility of nodes adjustment among each cache system.

When a content request emerges in ICN, the request is routed from the user to possible content providers. In order to simplify the strategy of request routing and content forwarding, most of caching mechanisms follow symmetric delivery in which contents are forwarded back along the reversed paths of requests routing paths. However, in SE mechanism, the content might be located either on/off the shortest delivery path. Therefore, content requests should not only be routed and looked up along the shortest path.
between users and original source. The nodes among SE caching system of the user are expected to be systematically requested to look up proper content copies. Systematic request routing and content forwarding strategy is designed for SE caching system as shown in Fig. 2.

The request routing and content forwarding strategy in SE starts from a request for content $i$ generated by receiver $R$. The request is firstly routed to edge node $E$ in level 1. If a copy of $i$ could be found in the $E$, $i$ is duplicated and forwarded back to $R$. If $i$ is not cached in $E$ at this moment, the request should be duplicated and broadcasted into all the semi-edge nodes from the level 2 of $R$'s caching system. Similar to the first level, a copy of content $i$ is forwarded back to $R$ if it could be found from any of semi-edge nodes. Otherwise, the content requests are routed from these semi-edge nodes to the original source of $i$ cross the core network. In this case, a copy of $i$ is forwarded back to the receiver from the source.

The request routing and content forwarding strategy of SE caching is derived from edge caching. Besides the benefits from caching point of view, the fixed maximum caching scale also reduce the latency by saving the looking up time during the request routing process.

3.2 Content Caching Strategy

Once a content is requested by a user, it is supposed to be cached in-network. Caching allocation strategies decide how many copies of this content should be cached and which caching node(s) among caching system are suitable to be selected for optimal caching efficiency. In SE caching system, two novel content allocation strategies, SE-U (Unicast Semi-Edge) and SE-B (Broadcast Semi-Edge), are proposed in order to adapt for different fog network topologies and scenario contexts.

3.2.1 SE-B: Broadcast Semi-Edge Caching

As shown in Fig. 2, after the request routing, content $i$ could only be found in three possible positions, edge node (a), semi-edge nodes (b) or original source (c). Under SE-B, if the content is found in edge node, it means the copy of content $i$ has already be cached in caching system lever 1. As level 1 is the nearest caching position to the receiver, the system requires no caching adjustment in this case. Else if the content is forwarded to receiver from any node among semi-edge nodes in level 2, it means there is no copy in edge node level. Similar to LCD strategy, the content copy is moved into level 1 edge node cache, which is one level down towards receiver. In the third case, if the content is provided by original source, it indicates that this content has not been cached in any nodes among the caching system of receiver. Meanwhile the content is forwarded back to receiver, the copies of the content are broadcasted (c1) and cached (c2) in all semi-edge nodes in cache system level 2.

SE-B is an online, off-path content allocation strategy. Similar to edge caching, SE caching only in network fog cuts downs the caching hardware and management cost in core network. In order to reduce the content redundancy in network, each content delivery session under this strategy only keeps one content along the delivery path. As an enhancement of edge caching, SE caching with hierarchical structure can increase the cache hit ratio of content allocation strategies. Because of the overlapped nodes belonging to multiple systems, popular contents copies cached in these nodes are able to satisfy more than one users.

3.2.2 SE-U: Unicast Semi-Edge Caching

SE-B introduced in Section 3.2.1 broadcasts content copies into level 2 caching nodes (c1 & c2) in the case of cache miss. However, the broadcast operation from content source to caching system may adversely increase the link load of core network. Moreover, the similarity of caching catalogue in semi-edge nodes increases the content redundancy as well. Thus, we refined the 2nd level caching from broadcast into unicast in SE-U.

When the request is satisfied by cached copy from caching system (a & b), caching operations under SE-U is as the same as the ones in SE-B. If the content cannot be found in any nodes among the caching system, a content copy is supposed to be cached in one selected node rather than all the semi-edge nodes. This unicast object node is selected according to the graph characteristics of nodes in level 2. Because all the nodes among semi-edge nodes have the same hop distance away from the receiver, the ideal content allocation principle is to select a node belonging to as many caching systems as possible. In this work, betweenness centrality is chosen as the graph-based characteristic to represent the caching weight of nodes. When the content is forwarded back from the receiver, the betweenness centrality value of every caching node in level 2 is calculated. One content copy is then forwarded (c1) and cached into the caching node with highest betweenness centrality (c2).

Under SE-U, each delivery session only leaves a single copy of requested content in the caching system of the user. Besides
the users connected to the same edge node, caching nodes can be shareable because nodes with higher centrality is more possible be included in more than one caching systems. Compared with the SE-B, SE-U is more suitable when the content popularity of requests is relatively concentrated.

4 FOG TOPOLOGY DESCRIPTION APPROACH

SE-U and SE-B are strongly coupled with the topology of network edge. Thus this chapter describes a novel approach to present the topological features of network fog. It aims to find out the suitable SE strategy (SE-U or SE-B) for each given topology. In order to clarify the definition of three proposed parameters, a simple topology is designed as an example shown in Fig.3. In this example topology, there are 1 source, 6 receivers (Receiver 1-6) and 7 forwarding hops with cache capacity (Node A-G).

- **Multi-User Degree**
  Multi-user degree is defined as the average number of caching systems allocated to per user. In the example topology, receiver 2 and 3 share the same edge node B. Thus the SE caching system of receiver 2 and 3 is the same one (consists of node A, B & C). Similarly, receiver 5 and 6 share the caching system with edge node G. Receiver 1 and 4 has their individual caching system respectively. In total, four Semi-Edge caching systems are shared by six receivers. multi-user degree is calculated by dividing the number of caching system by the number of receivers. So the multi-user degree in example topology is 0.67.

  According to this definition, topology with high multi-user degree is supposed to achieve better performance in SE caching system. Because high multi-user degree means a content copy in caching system can serve more users.

- **Edge Serve Ratio**
  Parameter edge serve ratio indicates the comparison performance of SE caching against edge caching. Node A is the edge node of receiver 1 meanwhile it is also the semi-edge node of receiver 2 and 3. In another word, Node A has three served receivers, which are one direct receiver (receiver 1) and two undirected receivers (receiver 2 &3). Similarly, nodes B-G also have specific number of served receivers which consist of direct receivers and/or undirected ones. When the fog topology of a network is given, its edge serve ratio is the percentage of the number of direct receiver over the number of all served receivers. In the example case, the sum number of direct receiver is 6 while the served receiver number is 15. Therefore, the edge serve ratio in example topology is 0.4.

  Under edge caching strategy, only caching nodes with direct receivers are involved in the caching process. Therefore, the definition of edge serve ratio can explore the performance differences between SE caching system and edge caching to some extent.

- **Caching System Density**
  Caching system density describes the overlapping degree of caching systems in network fog. For the reason that all the nodes A-G are within two hops away from end users, there are seven caching nodes in the example topology. Caching system density is defined as the quotient of the sum of served receivers divided by the sum of caching nodes. In the example topology, the sum of served receivers is 15 in total. So the value of caching system density is 2.143.

  In fog network topologies, the bigger Caching system density is, the more users are served by each caching node. It means caching systems in network is highly overlapped. Considering the proposed SE strategies, SE-B is relatively suitable to topologies with high caching system density while the SE-U is more suitable to low overlapped caching network.

5 SIMULATION EXPERIMENTS

Semi-Edge Caching mechanism is implemented and evaluated in ICN simulator Icarus [19]. The scenario configuration of experiments is described in Section 5.1. Section 5.2 compares our proposed SE-U and SE-B against edge caching strategies.

5.1 Scenarios Configuration

In order to compare caching strategies, the same content replacement policy, Least Recently Used (LRU), is applied to all the scenarios in our simulation. Moreover, parameters like content categorise size, numbers of warmup & measured requests are also fixed. Cache hit ratio and latency are used to represent the caching and network performance respectively. In order to avoid the unfairness caused by caching capacity sum in our simulation scenarios, the sum cache capacity of all the caching nodes in network is equal to the capacity of the single node in edge caching. Due to the increasing of involved caching nodes, the capacity of each in-network caching nodes in SE caching system is smaller compared with which in edge caching.

In simulator Icarus, each scenario is defined by a network topology and a request generator. For request generation, content requests emerge uniformly by the users following Poisson Arrival. The only variable parameter of request generator is content popularity coefficient in stationary Zipf’s distribution.

Regarding to network topology, three typical topologies GEANT, GARR and WIDE [14] are imported and configured for our experimental analysis.

5.2 Comparison Results

The performances of SE-B and SE-U strategies are compared with edge caching. According to the experimental results, the correlation between cache and network performance is explored. As the results shown in Fig.4, network latency is usually inversely related to cache hit ratio. Both of SE strategies can obtain lower latency and higher cache hit ratio than edge caching strategy. In concrete, the performance of SE-U is slightly better than SE-B.
Considering the comparison results in various topologies, Table 1 shows the scenarios when the popularity coefficient is set into 0.9 and cache size of network is the 1.5% of content catalogue. The performance of both SE-U and SE-B strategies are always better than edge caching in the three topologies. For example, the cache hit ratio of SE-U strategy increases by 48% than edge caching and the latency reduces by 7.6%. Concretely, the distinct suitability of SE-U and SE-B in different network topologies will be discussed in Section 6.1.

### Table 1: Comparison Results against Edge Caching

<table>
<thead>
<tr>
<th>Topology</th>
<th>Strategy</th>
<th>Latency</th>
<th>Cache Hit Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEANT</td>
<td>Edge Caching (baseline)</td>
<td>72.53</td>
<td>17.65%</td>
</tr>
<tr>
<td></td>
<td>SE-U</td>
<td>67.02</td>
<td>26.17%</td>
</tr>
<tr>
<td></td>
<td>SE-B</td>
<td>68.02</td>
<td>25.09%</td>
</tr>
<tr>
<td>WIDE</td>
<td>Edge Caching (baseline)</td>
<td>63.36</td>
<td>20.02%</td>
</tr>
<tr>
<td></td>
<td>SE-U</td>
<td>60.88</td>
<td>25.48%</td>
</tr>
<tr>
<td></td>
<td>SE-B</td>
<td>61.64</td>
<td>24.59%</td>
</tr>
<tr>
<td>GARR</td>
<td>Edge Caching (baseline)</td>
<td>69.53</td>
<td>15.49%</td>
</tr>
<tr>
<td></td>
<td>SE-U</td>
<td>69.25</td>
<td>17.93%</td>
</tr>
<tr>
<td></td>
<td>SE-B</td>
<td>68.13</td>
<td>19.59%</td>
</tr>
</tbody>
</table>

### 6 FURTHER EVALUATION

In this section, the effect of scenario factors on strategy performance are further explored. In concrete, Section 6.1 evaluates the performance results against topologies based on the description approach introduced in Chapter 4. Subsequently, SE caching system is further evaluated in Section 6.2 with the considerations of traffic patterns and cache capacity.

#### 6.1 Effects of Fog Topology

According to the definitions in Chapter 4, the fog part of GEANT, WIDE and GARR can be described as Table 2. Moreover, the simulation results of cache hit ratio against topologies are plotted in Fig.5, which could coincide to the fog topology analysis properly.

### Table 2: Topology Comparison by Description Parameters

<table>
<thead>
<tr>
<th></th>
<th>GEANT</th>
<th>WIDE</th>
<th>GARR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multi-User Degree</td>
<td>0.105</td>
<td>0.090</td>
<td>0.035</td>
</tr>
<tr>
<td>Edge Serve Ratio</td>
<td>27.6%</td>
<td>31.6%</td>
<td>25.3%</td>
</tr>
<tr>
<td>Caching System Density</td>
<td>1.813</td>
<td>2.714</td>
<td>4.15</td>
</tr>
</tbody>
</table>
of original source to some extent, the average waiting time of users are not benefits efficiently. With the increasing of network cache capacity, the latency performance becomes inversely related again. In the strategy design, SE system is possible to adjust network cache size corresponding to the content popularity context. Because the caching node amount of caching system could be configured flexibly without hardware modification.

7 CONCLUSIONS & FUTURE WORK

In this paper, a new generic caching mechanism SE was proposed for various content delivery services in different network types. SE mechanism allocates hierarchical caching system to each user. A request routing and content forwarding strategy was proposed within SE. In order to fit in different scenarios, SE-U and SE-B are proposed as two content allocation strategies. According to the empirical simulation studies, the content delivery performance of SE is obviously better than edge caching. The results show SE can increase cache hit ratio by 25%-45% depending on different scenarios. Meanwhile the latency always decreases along with the growth of cache hit ratio. Moreover, for SE strategy selection, an analysis approach of fog topology is explored and validated combined with the strategy performance results. SE strategies are also further evaluated under various network contexts in three typical network topologies. The effect of scenario contexts (e.g. caching capacity, content popularity) on content delivery performance is concretely analysed as well. In the future, the work will be extended with more comprehensive traffic model. Moreover, the quantitative analysis of hierarchical level will also be enhanced.

REFERENCES