A REVIEW OF CACHING STRATEGIES AND ITS CATEGORIZATIONS IN INFORMATION CENTRIC NETWORK

MOHAMMAD ALKHAZALEH¹, S. A. ALJUNID¹, NASEER SABRI¹

¹Advanced Communication Eng., Centre of Excellence, School of Computer and Communication Eng.
University Malaysia Perlis

ABSTRACT

Information-centric networking (ICN) is one of the rising Internet paradigms proposed to beat the defect of the current host-centric Internet. ICN relies on the name to access contents rather than its original location, which provides substantial flexibility for users to obtain the contents. In-network caching is one of the most important features of the ICN because it has a significant role in improving network performance such as reducing server load, congestion, and delay caused by users. In-network caching is managed by a caching strategy that determines what, where and when to cache the content, to make content is available for requesters without going to servers. In recent years, caching strategies have attracted considerable interest from researchers, and they have proposed many caching strategies to manage the contents to enhance the performance of ICN. In this paper, the caching strategies and its characteristics are clearly described and discussed; furthermore, the caching strategies are categorized based on a set of common characteristics to be easy to understand.

Keywords: Information-Centric Networking, Content Caching Strategies, In-network Caching, Caching Mechanisms.

1. INTRODUCTION

Due to the rapid growth of the Internet, the current P-2-P system has many drawbacks when it deals with a large scale of content distribution[1]. Nowadays, users are interested in accessing content other than their original location [2], especially a video that is widely used these days. For that, a new infrastructure named ICN has been proposed as an alternative to the current Internet.

The main idea about ICN is to access the content by the name other than the original location [3][4]. There are many architectures built on this idea and all fall under the ICN but differ in the implementation such as NDN[5], CCN[6], DONA[7].

ICN includes in-network caching[8], which is one of the main advantages of ICN. In-network caching improves network performance such as reducing server load, congestion, and delay caused by users [9] [10]. In-network caching allows to cache copies of content on routers in different locations, in order to make this content closer to the user. The caching strategy is responsible for managing the contents and selecting the appropriate location to cache a copy of content, and the performance of the in-network caching depends on a caching strategy, which consists of two parts, placement which chooses the appropriate location to cache a copy of the content, and replacement is responsible for evicting the content from the cache when it is full and new content has come to store in this cache [11].

In recent years, the researchers have been interested in ICN and especially the In-network caching, and have introduced several strategies for managing the contents in the Cache, but these caching strategies differ in term cache decision. For that, at the beginning of this review will be clarified the In-network caching feature in ICN and its importance. Then the categorizations of caching strategies will be defined, which based on some shared attributes between the principles of caching strategy to take decision to cache the content. After that, a comprehensive review of caching strategies will be introduced, which explain and summarize the existing caching strategies to be easy to understand. In addition to mentioning the modern caching strategies that have not been mentioned in the previous review papers such as MAGIC [12],
BidCache [13], LCLCS [14], APP [15], and PGBCS [16]. Then, the caching strategies will be classified based on these categorizations to realize and understand what are the similarities and differences between the principles of caching strategies to take decision to cache the content, which will be described in the following lessons:

2. CACHING IN ICN

ICN includes in-network caching features at every content router [8][11], which allow content to be cached at some of the content routers in the network, therefore provide efficient data delivery[17]. The caching consider one of the main advantages of the ICN, it allows to users and content routers to satisfy locally without going to servers, when consumer generate request for content, at any content router must cache the content to support other users and content routers maybe request the content and satisfy locally, this mechanism named placement, also which content to be replaced first when the cache is full, this mechanism named replacement [18][19].

Caching in ICN is an important thing because the efficiency of ICN based on caching strategies that manage the contents distribution along networks, which make the contents are available when the user requests these contents. The researchers proposed many caching strategies to manage the content distribution among the networks, but these caching strategies differ in term cache decision, some of these strategies based on the path, and other on probability, popularity, corporation, and content type.

3. CATEGORIES OF CACHING MECHANISMS

The objectives of caching mechanisms are to distribute content copies on the delivery paths, minimize content delivery time, and lessen the number of replicas. The current caching mechanisms are organized into several categories according to their attributes.

3.1 Path Caching:

Caching mechanisms can be divided into off-path caching and on-path caching[20]. For off-path, a Data packet is not necessarily to cache at the content router across the path to the requester. In contrast, a Data packet is cached across that path in on-path caching.

3.2 Redundancy Caching

Redundancy caching refers to a situation where a content is cached at more than one content router between requester and server (downloading path). Otherwise, it is a non-redundancy caching.

3.3 Cooperation Caching

Cooperative caching refers to the situation where content is cached according to the information provided by other content routers. It can be classified into cooperation implicit and explicit cooperation. With implicit cooperation, the delivery of content routers’ cache states through an extra advertisement mechanism is not necessary. Nonetheless, additional operations are needed to assist content routers in making forwarding or caching decision. In explicit cooperation, if a content router holds a copy of content, it has to advertise its cache state to other content routers.

3.4 Popularity Caching

The Popularity is the number of request for content at the content router, some of the caching strategies based on popularity to take the decision to cache content, and others don’t do that, for that, the caching strategies can be divided into Popularity Caching and Non-Popularity caching.

3.5 Distribution Caching

Some of caching strategies based on distribution, as the distance between the content router and the server or content router position. For that, caching strategies can be divided into Distribution Caching and Non-Distribution Caching.

3.6 Probability Caching

Some of caching strategies select content router to cache content based on content Probability, for that, caching strategies can be divided into Probability Caching and Non-Probability caching.

3.7 Content Type Caching

Every caching strategy has used Content type as, packet, chunk object, file and content, the packet and chunk consider same, also object and file and content consider same, for that, caching strategies based on Content type can be classified into chunk caching and content caching.

4. EXISTING CACHING MECHANISMS

This section explains the current caching mechanisms and the decision-making process on content caching.
4.1 Leave Copy Everywhere (LCE) [21].

For CCN/NDN, the default caching mechanism is LCE. LCE aims to cache the Data-packet identical to the Interest-packet sent by the requester in every content-router across the line from the content-router responding to the Interest packet to the requester, as shown in Figure 1.

![Figure 1: LCE operations.](image)

4.2 Leave Copy Down (LCD) [21].

The objective of LCD is to reduce the caching redundancy by including a caching proposal flag bit in the header of Data-packet. The flag bit will be activated when the content source or content router with a replica of conforming content receives a request. Using this bit, the content-routers across the transferring path will evaluate the necessity to cache the Data-packet.

In LCD, a copy of content will leave at the content router in one level down towards the requester after each request, as shown in Figure 2.

![Figure 2: LCD operations.](image)

4.3 Move Copy Down (MCD) [21].

The approach employed by MCD in minimizing the caching redundancy is similar to LCD. In MCD, the copy of content will move at the content router in one level down towards the requester after each request, as shown in Figure 3.

![Figure 3: MCD operations.](image)

4.4 Caching with Probability (Prob(p)) [21].

The rationale of caching with probability is to reduce caching redundancy and enhance efficiency. With probability \( p \), every content-router in Prob(p) will cache a Data-packet. However, for a likelihood of \( 1-p \), the Data-packet will not be stored. A random number in a range of zero to one will be produced once a Data-packet arrives at a content-router. The content-router will cache a copy of the Data-packet if the random number generated is less than \( p \). Otherwise, as demonstrated in Figure 4, the Data-packet will be forwarded without being cached.

![Figure 4: Prob(p) operations.](image)

4.5 Breadcrumbs [22].

Figure 5 shows the caching mechanisms of Breadcrumbs. Along the downloading path, a Data-packet is stored at every content-router. When the Data-packet is cached, a data routing history, also known as a caching-trail is created by a content-router. This trail includes content name, forwarding time, upstream, and downstream. Upstream indicates the origin of an incoming Data-packet at the content router, while downstream indicates the destination of an outgoing Data-packet from the content router.

For every caching trail, Breadcrumbs assign a threshold denoted as Tha. The next hops of Interest-packets will be decided based on the content name’s caching-trail at a content-router. When an Interest-packet is received by a content-router nonetheless a cached copy in its Content Source:

\[
\text{If } (\text{Interest-packet time} - \text{Forwarding Time}) < \text{Tha} \quad \text{Then} \quad \text{downstream} \\
\text{Else} \quad \text{upstream}. 
\]

![Figure 5: Breadcrumbs operations.](image)

4.6 Probcache [23].

In Probcache, the probabilistic caching mechanism includes a Time-Since-Inception (TSI) part in the header of Interest-packet and a Time-
Since-Birth part in the header of Data-packet. Initially, the TSI cost is at zero once an Interest-packet is sent out by a user. After each hop, the value will increase by one. On the other hand, the TSB value is set at zero initially as soon as a Data-packet is sent out by a content source. After that, the value will increase by one per hop. Nonetheless, the final value of TSI remains unchanged.

TimesIn \((T_{\text{In}})\) refers to the sum of times that a packet can be saved on a path. Using TimesIn, the available caches on the downloading path can be estimated.

TSI value contained in the Interest-packet field header and TSB cost contained in the header of a Data-packet are denoted by \(y\) and \(z\), respectively.

\[ T_{\text{Tw}} \text{ value refers to the duration that a content must be stored on a distribution pathway such as 10.} \]

For the distribution pathway, the cache size of content-router \(n\), \(N_i\), is represented by the duration of cached content maintained in a cache. The mean of the cache size of a distribution pathway, \(N_z\), is determined by the cache size of the content-routers on a delivery path.

\[
T_{\text{In}}(z) = \sum_{i=1}^{y} x_i N_i \\
\text{CacheWeight}(z) = \frac{z^y}{y} \tag{2}
\]

\[\text{ProbCache}(z) = T_{\text{In}}(z) \times \text{CacheWeight}(z) \tag{3}\]

ProbCache calculates the TimesIn and the cache weight, then calculates caching probability \((\text{ProbCache}(z))\) at each content router by using equation 3, a content router with a big \((\text{ProbCache}(z))\), compared to the content routers along the path, will be caching contents with higher probability as shown in Figure 6.

The sum of content chunks to be stored by a content-router is strongminded by the Chunk Marking Window \((\text{CMW})\), the CMW is determined by CMW state \((n)\) and CMW base \((x)\), Thus, CMW ranges will be from \(\sum_{j=0}^{n-1} x^j + 1\) to \(\sum_{j=0}^{n} x^j\). The WAVE operations are described in Figure 8.
4.9 Cache-Aware Target Identification (CATT) [25].

The objectives of CATT are robustness, availability, diversity, and adaptability. A Data-packet is stored at a single content-router by CATT across the downloading path in every network. With the introduction of expected quality of content, CATT assists content-routers in forwarding Interest-packets. As shown in Figure 9, once a content-router stores a content, the expected quality of such router is denoted as Qc. Otherwise, the expected quality of the content at such router will be denoted as Qua(Ri).

\[
Qua(Ri) = \frac{Qc}{\text{Hop}_r + 1}
\]

(7)

The sum of hops from one router source to the base server as well as the content-router that caches the content is denoted as Hopr.

Instead of FIB, CATT employs a Potential Based Routing (PBR) for content retrieval. In CS, when there is no cached replica of the content demanded by Interest-packet, the content-router that receives the packet would examine the expected quality of content of its neighbors’ PBR. Then, the content-router would select the neighbor with greatest expected quality of content to forward the Interest-packet.

![Figure 9: CATT operations.](image)

4.10 A Chunk Caching Location and Searching Scheme (CLS) [26].

As an advanced version of MCD, CLS seeks to improve file download time and server workload. Hence, a maximum of one replica of a chunk is stored by CLS on the pathway amid a server and a requester. When there is a request, the copy of chunk will move one level down towards the requester. In the case of cache eviction, the chunk will move one level up towards the server. In CLS, a content-router will maintain a caching table that consists of several entries and each of the entries is a 4-tuple, namely the content-name (Name), the upstream content-router (Content-routerIDin), the downstream content-router (Content-routerIDout), and the number of hop to the content source (Hop).

As shown in Figure 10, in a caching trail, the value of Hop field would specify the caching position of the consistent content.

![Figure 10: CLS operations.](image)

When a content-router obtains an Interest-packet with content less in the CS, the content router can use the Hop number in the caching pathway to determine the name that is able to be employed to identify the subsequent hop of the Interest-packet, either the upstream or the downstream router.

The network administrator would determine the threshold value (Tha) in advance.

If \( Tha \leq \text{value of hop} \) then downstream
else upstream.

4.11 Edge Caching [27].

As shown in Figure 11, edge caching allocates content nearer to its users by storing content in the latest content-router on a distribution pathway. The objective here is to decrease the number of hops needed to reach a content source. This in turn leads to the reductions of both time used for content delivery and traffic within a network caused by the transmission of content requests.

![Figure 11: Edge Caching operations.](image)

4.12 Distributed Cache Management (DCM) [28].

The features of DCM are the maximization of traffic volume served from the caches and the minimization of the bandwidth cost. According to DCM, a Cache Manager (CM) is installed so that its cache state is exchanged with other CMs. The decision about the replacement of cached content depends on the distributed on-line cache management algorithm employed by the CM.
Let \( G_n = \langle V, E \rangle \), where \( V \) is content-routers, and \( E \) is links between content-routers. For a better understanding, the main algorithm parameters are as below:

- \( S \) is the whole sum of saved content of an assumed network \( G_n \).
- \( r^s R_v \) is the inward demand ratio at \( R_v (R_v \in V) \) based content \( s (s \in S) \).
- \( hsR_v \) is true once \( s \) is saved by \( R_v \).
- \( hsR_v = \) if \( (s \) is saved by \( R_v) \) Then 1 else 0.

DCM aims to obtain an applicable configuration in minimizing the cost of network traffic. A linear programming is formulated as below:

\[
T(H) = \sum_{i=1}^{s} r^s R_v dR_vR_m \quad (8)
\]

Where \( R_m \) represents the nearest cache-router that stores a replica of a given content \( s \) and \( dR_vR_m \) represents the number of hops between \( R_y \) and \( R_m \).

The first constraint indicates that every a stored content is cached in at least one content-router. The additional restriction is the limited caching capacity for each content-router.

The myopic cache management algorithm is employed and explained as below:

- First stage: \( \text{CM}(R_v) \) chooses a content such as content \( s \) and determines if there is a locally cached content \( X \).
- Second stage: If content \( s \) is yet to be cached by \( R_v \), the traffic gain, \( g_s = T(H_i) - T(H^*) \), will be estimated by \( \text{CM}(R_v) \) when it is saved by \( R_l \), later, which \( H^* \) represents that \( hsR_v \) is equal to true and the other cache formations persist constant. If content \( s \) has been cached, the traffic loss, \( l_s = T(H^*) - T(H_i) \), will be estimated by \( \text{CM}(R_v) \) when \( R_v \) evicts the content \( s \), where \( H^* \) represents that \( hsR_v \) is equal to untrue while supplementary cache formations persist constant.
- Third stage: \( \text{CM}(R_v) \) repeats the first and second steps until the cached contents are completely chosen. The highest existing gain content is regarded as a saved content candidate (for example \( i \)) while the content of a lowest traffic lost is considered a replacement applicant (for example \( j \)). \( \text{CM}(R_v) \) will fetch the content \( i \) and evict the stored content \( j \) when the relative gain, \( R_g = g_i - l_j \), is positive. \( hiR_n \) and \( hjR_v \) are simultaneously set to be true and false, respectively.
- Fourth stage: The advertisement of new cache state is made by \( \text{CM}(R_v) \) to other CMs.
- Fifth stage: Repeat each step in other CMs until there is no extra beneficial replacement.

4.13 Popularity based Caching Strategy for Content Centric Networks (MPC) [29].

Instead of distributing widespread content on the way to the neighbors on a delivery path, the key focus of MPC is the distribution of popular content towards all neighboring content-routers. Under MPC, only popular content is cached. Each content-router in the network can store two types of information, namely “content name” and “popularity count”, hooked on a special table known as popularity table.

Figure 12 shows that every content-router across the downloading path sums the incoming needs for every content name locally and the data is stored in a popularity table.

If Content Popularity \( \geq \) Threshold, Then Cache & Popular, Else Unpopular.

Given a cached popular content, the particular content-router would send a message to its neighboring content-routers and recommend them to store the popular content. After that, the decision to cache the said content is made by the neighboring content-routers based on their respective caching policies.

4.14 Intra-AS Cache Cooperation [30].

By allowing the public visibility of cached content, Intra-AS Co can reduce the caching redundancy.

In Intra-AS Co, each content-router maintains two types of cache brief list, namely Exchange Cache-Summary Table (ECST) and Local-Cache Summary Table (LCST).

The ECST stores the information of cached contents at a router’s neighboring content-routers.
while the LCST stores the information of the locally cached content chunks.

Furthermore, each content-router regularly advertises its summaries of local-cache to its directly linked neighbors. A content-router will store the information of router identifiers and content names into its ECST when summaries are received from its directly linked neighbors.

**Figure 13: Intra-AS operations.**

Once an Interest-packet reaches at a content-router, its cached copy will be searched in the router’s CS. If there is no match found, the content-router will use its ECST to check on the availability of the desired content chunk cached in its directly linked neighbors. Based on the router’s FIB, the Interest-packet will be promoted to the following hop when no cached replica is found, as shown in Figure 13.

### 4.15 Centrality-based Caching (CBC) [31]

The objectives of CBC are to reduce the latency of content delivery, alleviate traffic and congestion given that a cache hit will result in lesser paths traversed by a content, and reduce server load since one cache hit leads to one less request.

\[ BC(R) = \sum \frac{n(R_i,R_j,R_y)}{n(R_i,R_j)} \]  

(9)

\( n(R_i,R_j,R_y) \) is the numeral of shortest paths from \( R_i \) to \( R_j \) through \( R_y \). While, \( n(R_i,R_j) \) represents the number of shortest paths amid \( R_i \) and \( R_j \).

As shown in Figure 14, Data-packet is stored at the “important” content-router with the greatest betweenness centrality (\( BC(R) \)) rate across the delivery path.

**Figure 14: CBC operations.**

For CBC, the header of Interest-packet includes a part that records the maximum \( BC(R) \) rate.

When an Interest-packet is promoted, the \( BC(R) \) values of the forwarding content-router and the one stored in the Interest-packet are associated. If the router’s \( BC(R) \) rate is higher than the individual stored in the Interest-packet, the router will replace the \( BC(R) \) value of the Interest-packet with its own \( BC(R) \) value.

Similar to Interest-packet, the \( BC(R) \) values of the receiving content-router and the one stored in Data-packet are compared. If the values are the same, the content-router will cache the Data-packet into its CS.

### 4.16 Two Layers Cooperative Caching (TLCC) [32]

For TLCC, there are several groups of clustered CCN routers within one Autonomous System (AS). Using consistent hashing, each group collaborates in storing contents and eliminating duplicate contents.

One AS is made up of two groups, as shown in Figure 15, namely Upper-Layer Group (ULG) and Lower-Layer Group (LLG). ULG is linked to LLG or further AS, this connection will lead to the storage of chunks corresponding to the keys of respective router, whereas LLG is connected with users, this connection will lead to the storage of popular chunks based on keys and the popularity threshold value.

Two types of caching algorithms are employed in TLCC, namely Cache Decision for Upper Layer Group (CDAULG) and Cache Decision for Lower Layer Group (CDALLG).

**CDAULG:** When ULG routers receive a chunk, the chunk name is checked and hashed to identify the corresponding router. ULG will cache the chunks if its hashed value is within the key range of current router. Otherwise, the chunk is then forwarded to the dedicated router.

**CDALLG:** When LLG routers receive the Data chunk, the routers examine the chunk name and Local Popularity Count (LPC). LLG will cache the
chunks when there is an above-threshold LPC and the hashed value is within the key range of the current router. Otherwise, the current router will just forward the chunk to the client or other routers through requested face(s).

![Figure 15: The groups of Intra-AS](Image)

### 4.17 A Distributed MAX-Gain In-network Caching Strategy in ICN [12]

The objectives of MAGIC are to reduce bandwidth consumption with a joint consideration of hop reduction and content popularity. Besides, cache replacement penalty is considered in the decision-making process for cache assignment to decrease the caching operations.

Let $G = \langle V,E \rangle$, where $V$ is content-routers, and $E$ is associations between content-routers. For a better understanding, the main algorithm parameters are as below:

- $M$: The requested content where $M = \{m_1, ...., m_{|M|}\}$.
- $r^m_v$: The request rate (calculate as the sum of Interest-packet per second) at content-router $v$ for content $m$.
- $h^m_v$: The number of hops from content-router $v$ to the base server that preserve content $m$.
- $C_v$: The content saved at content-router $v$ where $C_v \subseteq M$.
- $c_i$: The cache capacity of content-router $v_i$.

First, MAGIC calculates the potential place gain of cache content $m$ at router $v$; second, it calculates the cache replacement penalty, then Local Gain of cache content $m$ at router $v$.

$$\text{PlaceGain}^m_v = r^m_v \times h^m_v$$

$$\text{RePlacePenalty}_v = \begin{cases} \min \{ m \in C_v \mid r^m_v \times h^m_v \}, & \text{if } |C_v| = c_i \\ 0, & \text{if } |C_v| < c_i \end{cases}$$

LocalGain$^m_v = \text{PlaceGain}^m_v - \text{RePlacePenalty}_v$ (12)

In MAGIC, if the drawback of substituting a saved content is greater than the gain of caching a new content, then there will be neither cached content eviction nor cache replacement.

![Figure 16: MAGIC operations](Image)

In order to identify the content-router that has a greatest $LocalGain$ for content $m$, an Interest-packet is attached with the maximum local cache-gain across the pathway in the latest denoted by $MaxGain$ field. An Interest-packet with a zero $MaxGain$ cost in the header is sent to request content $m$.

When the Interest-packet is received, the value of $LocalGain^m_v$ is estimated by each router $v$ and the estimated cost is likened with the value contained in the $MaxGain$ field.

When the recorded $MaxGain$ value is lower than the local cache-gain of content-router $v$, the $MaxGain$ cost in the Interest-packet will not be updated. When the Interest-packet arrives at a content-router or the initial server, the $MaxGain$ part in the Interest-packet equals to the extreme local cache-gain across delivery path.

The Data-packet is attached with $MaxGain$ value of the Interest-packet, and then the Data-packet will send to node that request and cache at content-router with the equivalent value as $MaxGain$ value in the Data-packet. Figure 16 shows the fundamental MAGIC’s operations.

### 4.18 In network Caching for ICN with Partitioning and Hash-Routing (CPHR) [33]

The objective of CPHR is to fully utilize ICN’s built-in caching capability. Hence, a new field known as $Egress Router$ is included in each FIB entry of a content-router in CPHR. This is to keep track the movement of Interest-packet for the content name from a field to a content source. Also, an entrance router includes two extra parts in the header of Interest-packet.
Cache-content Name - To document the position of cached content-chunk for the content-name in the Interest-packet.

Egress-content Router - To show the name of outlet router conforming to the content-name in the Interest-packet.

Besides, content names are mapped to content-routers in CPHR via consistent hashing. Agreeing to the content-name hashed, the content-router will cache the content chunk for the content name hashed. Likewise, a content-router will hash the content name given by the Interest-packet and the Interest-packet is then forwarded to the content-router which the content-name is hashed to, as shown in Figure 17. This is done with the assistance of Cache-Name part in the Interest-packet.

The request will be satisfied when the content chunk is cached locally by its corresponding content-router. Otherwise, the Interest-packets will be forwarded to the content source with the assistance of the Outlet Router filed in the Interest-packet.

To attain this objective, CINC includes dual new lists in every content-router, namely the Collaborative-Router Table (CRT) and the Collaborative-Content Store (CCS). CRT aims to document the information about reachability of other content-routers within a network, while CCS aims to document the content names cached by other content-routers.

4.19 Cooperative in Network Caching (CINC) [34].

CINC is designed as a supportive caching technique for huge audiovisual streams such as online TV services. Using a customary of organized CRs, CINC aims to reduce the number of inquiries for time shifted TV that are responded by servers beyond the Internet Service Provider network.

In CINC, q is the number of content-routers, \( L_r \) is label of a content-router, and \( S_d \) is the order number of a Data-packet.

If \( L_r = S_d \mod q \), then the strategy cache copy at content router, else forwards, to avoid caching redundancy.

4.20 Auction-based In-network Caching in ICN (BidCache) [13].

The key idea of BidCache is that the right of caching is determined via bidding process. The Interest message will then piggyback the winning bid value. By comparing the present bid cost in the Interest-message and the local-bid that can be issued by the cache, each caching content-router will determine their bidding decisions. Prior to the forwarding of Interest message to the next-hop, the present uppermost bid value in the Interest-message is replaced with the bid cost of the content-router that decides to bid.

The information about the winner in the auction will be distributed to the participating caching content-routers using piggy-backing in the return Data-message.

1) Issuing a bid

The bid value contains a weighted \((w_1 - w_k)\) sum of local and non-local factors that are comparative to the requested content. Each parameter has distinct weight and this weight would influence the caching probability at a specific content-router. For example, a greater weight in a central content-router would lead to higher probability to cache popular content.

The weights can be influenced by the nature of content-router, network features and practice, or the adjustment made by the network manager. An example of bid construction is shown below:

\[
BID = < \text{local factors}> + <\text{context} - \text{request}>
\]
Motivating local factors to study for the bid cost could contain:

- The value of Least Recently Used (LRU)
- The value of Least Frequently Used (LFU)
- Link Bandwidth (BW)
- Cache size (CS)
- Hop Count (HOP)
- Link Delay (LD)

The bid cost is computed as follows:

\[
BID = w_1 \times f_1(LRU) + w_2 \times f_2(LFU) + w_3 \times f_3(BW) + w_4 \times f_4(CS) + w_5 \times f_5(HOP) + w_6 \times f_6(LD)
\]

Functions \(f_1, f_2,\) etc. allow the normalization of each parameter value for all caching content-routers. This is to ensure that the bids from content-routers with possibly huge dissimilar sceneries or abilities are adjusted with a similar scale.

On the other hand, the weights \(w_1, w_2,\) etc. allocate comparative importance to definite factors which are highly-prioritized based on the opinion of content-router. For example, a large weight is assigned to the certain content-routers with small CS to ensure the CS “dominates” the last bid cost.

Algorithm 1 demonstrates the process after a bid is issued by a caching content-router. Initially, an Interest-message is sent with \(r.bid\_winset\) to an indeterminate value. Then, the cache might place a bid if its estimated bid is equivalent to the bid transported in the Interest-message and this may lead to the previous bid value of the cache be overridden. Otherwise, the content-router will not bid.

Algorithm 1: Allotting a bid

\[
\begin{align*}
r & = \text{receive(Interest)} \\
c & = r.request\_context \\
l & = \text{local\_parameters} \\
b & = \text{issueBid}(l,c) \\
\text{if } b & > r.bid\_win \text{ or } r.bid\_win = \text{"undefined" then:} \\
r.bid\_win & = b \\
r.bid\_winset & = \text{True} \\
\text{else:} \\
r.bid\_winset & = \text{False} \\
r.request\_context & = \text{calculateRequestContext()} \\
\text{forward}(r)
\end{align*}
\]

2) Deciding the auction winner

When the producer or the holder of the duplicated requested data item receives the Interest message, the data item is returned across the Interest path. Moreover, the Data message will carry the information concerning the auction winner. As demonstrated by Algorithm 2, each content-router on the path will use this information to determine if it has the winning bid.

In the algorithm, an assumption is made where the Data message sent by the producer or the holder of duplicated requested data-item set the \(r.can\_cache\) part as True. When the content-router with the winning bid obtains the Data- message, the data-item is stored and the \(r.can\_cache\) part is set as non-true. This is to inform the following content-router that the bid has been won by previous content-router, thus the item does not need to be cached.

Algorithm 2: Determining the auction victor

\[
\begin{align*}
r & = \text{receive(Data)} \\
can\_cache & = r.can\_cache \\
\text{if } can\_cache & = \text{True then:} \\
\text{if } bid\_issued & = \text{True:} \\
\text{cache(Data)} \\
r.can\_cache & = \text{False} \\
\text{forward}(r)
\end{align*}
\]

Figure 18 shows the process of an auction.

Figure 19: BidCache operations.

4.21 Link Congestion and Lifetime Based In-Networking Caching Schemes (LCLCS) [14].

With the consideration of betweenness centrality, link congestion, and content popularity, LCLCS aims to sustain acceptable cache rate and significantly optimize network delay.

The process of LCLCS begins with an network analysis that leads to the identification of “important” content-routers in the content delivery path. After that, LCLCS applies the calculated lifetime to decide the eviction of current cached content. This is to ensure sufficient space for the incoming content to be cached.
In the centrality of content-routers within a network, the betweenness centrality is the key metric as shown below:

$$C_B(v) = \sum_{i \neq v \neq j \in V} \frac{S_{i,j}(v)}{S_{i,j}}$$  (14)

Where $S_{i,j}(v)$ is the sum of pathways in the range of $i$ to $j$ over content-router $v$, and $S_{i,j}$ is the sum of pathways from $i$ to $j$. Using the central management system with an assumption of known content-router’s betweenness value, the delivery route can be estimated offline or online.

Delay saving, $S_{T_{r,v}}$ refers to the time difference between obtaining content from the initial server and from the content-router $v$ (local-cache). While $L_{D}$ refers to a set of connections bridging the cache to the downstream requesters, $L_{U}$ refers to a set of connections bridging the upstream source to the cache. For every connection $i$ of $L_{D}$ and $L_{U}$, the set of available link capacities in $L_{D}$ is defined as $B_{LD}$ and the set of available link capacities in $L_{U}$ is defined as $B_{LU}$. Given that greatest bottleneck connections decide the entire throughput of the data flow, the throughput from the cache to the requester is the min $(B_{LD})$ and the throughput of the both ends in the pathway is the min $(B_{LU})$. Then, the approximation of postponement saving via content saving on content-router is shown as below:

$$S_{T_{r,v}} = \frac{S}{\min (B_{LD} \cup B_{LU})} - \frac{S}{\min (B_{LD})}$$  (15)

The content size is denoted as $S$. High $S_{T_{r,v}}$ value indicates longer duration is needed for contents to be cached in local. The purpose of this is to decrease the content download suspension.

Popularity shows the probability of content requested over a specific future timeframe. Given $R$ contents with associated popularity, the weights of popularity are estimated as below:

$$weight_r = \frac{p_r}{\sum_{r=1}^{R} p_r}$$  (16)

As the fundamental concept of LCLCS, lifetime determines the content storage and content deletion. Each content has a lifetime and its replica obtains the content’s lifetime before being included to the cache. When the content-router wants to cache a content but there is insufficient location, the content of a smallest lifetime is eliminated. According to this information, the lifetime of cache can be represented as below:

$$\text{Lifetime}_{r} = C_{L}(ST_{r,v} \cdot weight_r) \cdot C_2 \cdot \frac{S_{size}}{\text{base_time}}$$  (17)

Given constant $C_1$ and $C_2$, size is content size, and base_time is the basel cache time. The reason of including the content size is to enhance cache hit rate by caching compact object instead of the huge one. Given the dynamic nature of network state and content popularity, the concept of lifetime with diminishing nature is selected to ensure the feasible eviction of content that is difficult to obtain or popular from a cache.

When a content request is initiated, the relatively “important” betweenness set is documented by the request packet across the distribution path. The header of request packet will include the set and the field of sets is given as [base_set, max_set]. When the content distribution begins, the initial magnitude of set is given as base_set. For each three hops, there will be a unit increase in the size of set, equal to the greatest value of max_set. The size of set will decide the number of “important” content-routers. Prior to cache hit, the betweenness value of all intermediate content-routers will be compared and replaced with the betweenness value of $C_{BN}$ and the set will be constantly updated across the delivery path. Hence, $C_{BN}$ will carry the most important betweenness of content-routers across the path. While the data are transmitted at the hit content-router, the value of $C_{BN}$ is reproduced in the Data-packet. En route to the requester, the betweenness of each content-router is compared with the attached $C_{BN}$. Then, caching will happen if there is a match and sufficient space for the object. If there is a match but space is insufficient, expired contents will be removed so that content-router can cache the new object. However, the Data-packet will be promoted to the following hop when no match or no expired contents when there is insufficient space.
4.22 Adaptive Prioritized Probabilistic Caching Algorithm for Content Centric Networks [15].

Using caching probability as a foundation, the adaptive prioritized probabilistic caching algorithm aims to provide solutions for the limitations of previous caching algorithms. The technique emphasizes on content popularity and multiple priorities of Data-packets. Nonetheless, the caching probability considers the priority of data layers in its estimation.

Since certain Data-packets are more important, the priorities of Data-packets are divided into layers. The levels of importance for data layers are quantified and shown in Table 1.

Table 1: Data-Layers significance values

<table>
<thead>
<tr>
<th>Layers</th>
<th>Base-Layer (i=0)</th>
<th>Improved-Layer (1)</th>
<th>Improved-Layer (2)</th>
<th>Improved-Layer (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ_i</td>
<td>1.0</td>
<td>0.3</td>
<td>0.2</td>
<td>0.1</td>
</tr>
</tbody>
</table>

A random real number between [0,1] is generated when data arrive at CCN router. The generated number is then compared with the caching probability. If the caching probability is greater than the generated number, the Data-packet will not be saved in the CS, as shown in Figure 20. The types of calculation component include factors and events. The factors include additional factor (ac), caching probability (P), priority values (γ), and multiplicative factor (mc). According to events such as cache-miss and cache-hit, CCN router can adjust the caching probability in its estimation process. Cache-miss refers to a situation where the incoming Interest-packets are not fulfilled by the cached Data-packets. Since the requesters do not need the current contents in the CS, CCN router will cache the subsequent incoming high probability content. On the other hand, cache-hit refers to the situation where the incoming Interest-packets are satisfied by the cached Data-packets. With current high cache-hit, CCN router will decrease the probability of subsequent received Interest-packets. Since data are divided into multiple layers, each layer has its own caching probability. Given n layers of data, CCN-router will estimate and save caching probabilities of a value of n. Eq. (18) estimates iP at time t

\[
P_i(t + S) = \begin{cases} 1 & \text{if } P_i(t) \text{ is greater than 1} \\ P_i(t) \ast (c_m \ast \gamma_i), \text{hit is occured} & \text{if } P_i(t) + (c_a \ast \gamma_i), \text{Elsewhere} \end{cases} \tag{18} \]

As shown in Eq. (18), the caching probability at time t+S is estimated based on the preceding caching probability at time t. Once an Interest-packet reaches at CCN router, its consistent Data-packet is searched in the CS and the caching probability will be updated contingent upon a cache event. In the case of cache-miss, the multiplication of the supplementary issue and the priority value of data-layer is added into the caching probability. In the case of cache-hit, the product of multiplication factor and priority value is multiplied with caching probability and this leads to lower cache probability.

The decision of caching Data-packets at CCN routers is assessed depending the importance of the received Data-packet and the caching likelihood of received Data-packet layer. For instance, given that the inward Data-packet at time t+S is the base-layer, the caching decision is made through comparison of random real number and caching probability \( P_0(t + S) \).

4.23 Popularity and gain based caching scheme (PGBCS) [16].

The existing caching strategies lack choices in content placement and balanced distribution, thus, which lead to the problem of high user access time delay and low cache hit ratio. To solve the problem, PGBCS is proposed. The used symbols in PGBCS are defined in Table 2.

Table 2: Symbols for PGBCS

<table>
<thead>
<tr>
<th>Notation</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>o</td>
<td>Content chunk</td>
</tr>
<tr>
<td>vi</td>
<td>The identify of i content router</td>
</tr>
<tr>
<td>w_i</td>
<td>The average value of content chunks cached in vi</td>
</tr>
<tr>
<td>S_i</td>
<td>The occupied space of vi</td>
</tr>
<tr>
<td>r_i</td>
<td>The space of vi expected to take up</td>
</tr>
<tr>
<td>F_i</td>
<td>The free space of vi</td>
</tr>
<tr>
<td>C_i</td>
<td>The overall cache space of vi</td>
</tr>
<tr>
<td>d_i</td>
<td>The distance between vi and user</td>
</tr>
<tr>
<td>D</td>
<td>The total length of Interest path</td>
</tr>
<tr>
<td>q_i</td>
<td>The ratio of d_i and D</td>
</tr>
<tr>
<td>M (vi)</td>
<td>Caching gain of vi</td>
</tr>
</tbody>
</table>

There are four steps to calculate caching gain as following:
Step 1: calculating the average value of the content chunks cached in vi.
\[ w_i = \frac{1}{n} \sum_{j=1}^{n} (O_j) \]  
(19)

Step 2: Computing the utilization rate of vi.
\[ p_i = \frac{S_{i+r_i}}{c_i} \]  
(20)

Step 3: The distance factor.
\[ q_i = \frac{d_i}{D} \]  
(21)

Step 4: The calculation formula of vi caching gain index.
\[ M(v_i) = \frac{1}{w_i} \cdot \frac{1}{p_i} \cdot \frac{1}{q_i} \]  
(22)

When interest comes to the content router, check if the content router has free space or not, then it calculates the caching gain. When the cache hit, if content routers have free space on the download path, PGBCS will cache the chunk on the content router that has free space and is closest to the user as shown in Figure 22, if not, the strategy will cache the content on the content router with maximum caching gain as shown in Figure 23. When the content arrives to cache at the content router, PGBCS checks if the content chunk popularity value is not less than the lowest content chunk popularity value in the content router, the strategy will evict the content chunk that has the lowest value to cache the new content chunk, if no, the content chunk will be discarded.

In Path caching, the strategies are categorized into On-Path and Off-Path. The On-Path is used to clarify if the caching strategy caches a copy or copies of the content along downloading path between the server and requester such as LCE, LCD, MCD, Prob(p), Breadcrumbs, Probcache, Probcache+, WAVE, CATT, CLS, MPC, Intra-AS, CBC, TLCC, MAGIC, CPHR, Edge, BidCache, PGBCS, APP, and LCLCS. The Off-Path is used to clarify if the caching strategy caches a copy or copies of the content out the downloading path between the server and requester such as CINC and DCM.

In redundancy caching, the caching strategies are categorized into redundancy and non-redundancy caching. The redundancy caching is used to clarify if the caching strategy caches more than one copy of the content along downloading path between the server and requester such as LCE, LCD, Breadcrumbs, and Intra-AS, while the non-redundancy caching is used to clarify if the caching strategy caches one copy of the content along downloading path between the server and requester such as MCD, Prob(p), Probcache, Probcache+, WAVE, CATT, CLS, MPC, DCM, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, APP, and LCLCS.

In popularity caching, the caching strategies are categorized into popularity and non-popularity caching. The popularity caching is used to clarify if the caching strategy based on the number of requests for content to select a content router to cache the content, such as LCD, MCD, WAVE, CLS, MPC, DCM, MAGIC, PGBCS, and LCLCS. Otherwise, the caching strategy considers as non-popularity caching such as LCE, Prob(p), Breadcrumbs, Probcache, Probcache+, CATT, DCM, Intra-AS, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, APP, and LCLCS.

Generally, the Cooperative caching is used to clarify if the caching strategy caches the content based on information from other content routers. In Cooperative caching, the caching strategies are categorized into Implicit Cooperation Caching, explicit Cooperation Caching and non-Cooperation Caching. If the content router advertises information about what it has contents, in this case, the caching strategy considers as explicit Cooperation Caching, such as CATT, DCM, Intra-AS, TLCC, CPHR, and CINC. While if the content router advertises other information such as the number of requests for content or number of hops from the requester, in this case, the caching strategy considers as Implicit Cooperation Caching.

5. SUMMARY OF CATEGORIES OF CACHING MECHANISMS

After introducing the caching strategies and define the categorizations of caching strategies, in this lesson, the caching strategies will be classified based on these categorizations to realize and understand what are the similarities and differences between the principles of caching strategies to take a decision to cache the content.
Cooperation Caching such as LCD, MCD, Breadcrumbs, Probcache, Probcache+, WAVE, CLS, MPC, CBC, MAGIC, Edge, BidCache, PGBCS, and LCLCS. Otherwise, the caching strategy considers as non-Cooperation Caching such as LCE, Prob(p), and APP.

In probability caching, the strategies are categorized into probability and non-probability caching. The probability caching is used to clarify if the caching strategy caches the content based on probability such as Prob(p), Probcache, Probcache+, and APP. Otherwise, the caching strategy considers as non-probability caching such as LCE, LCD, MCD, Breadcrumbs, Probcache, Probcache+, WAVE, CLS, MPC, CBC, MAGIC, Edge, BidCache, PGBCS, and LCLCS.

Generally, the distribution caching is used to clarify if the caching strategy selects content router to cache the content based on content routers position such as the distance between content router and server, the distance between the content router and requester, edge router, neighbor router, and center router. In the distribution caching, the caching strategies are categorized into distribution Caching, and non-distribution Caching. According to that, LCD, MCD, Probcache, Probcache+, WAVE, CATT, CLS, DCM, Intra-AS, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, and LCLCS consider as distribution Caching. While LCE, Prob(p), Breadcrumbs, MPC, and APP consider as non-distribution Caching.

In content-type caching, the strategies are categorized into chunk caching and content caching. The chunk caching means the caching strategy divides the content into chunks, and each chunk needs interest message such as Probcache, Probcache+, WAVE, CLS, MPC, Intra-AS, TLCC, CINC, PGBCS, and APP. Otherwise, the caching strategy considers as content caching such as LCD, MCD, Prob(p), Breadcrumbs, CATT, DCM, CBC, MAGIC, CPHR, Edge, BidCache, and LCLCS.

The review of the latest caching strategies and its categorizations is summarized in Table 3.

<table>
<thead>
<tr>
<th>Category</th>
<th>Caching Mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Implicit Cooperation Caching</td>
<td>LCD, MCD, Breadcrumbs, Probcache, Probcache+, WAVE, CLS, MPC, CBC, MAGIC, Edge, BidCache, PGBCS, LCLCS.</td>
</tr>
<tr>
<td>Explicit Cooperation Caching</td>
<td>CATT, DCM, Intra-AS, TLCC, CPHR, CINC.</td>
</tr>
<tr>
<td>Non-cooperation Caching</td>
<td>LCE, Prob(p), APP.</td>
</tr>
<tr>
<td>On-path Caching</td>
<td>LCE, LCD, MCD, Prob(p), Breadcrumbs, Probcache, Probcache+, WAVE, CATT, CLS, MPC, Intra-AS, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, APP, LCLCS.</td>
</tr>
<tr>
<td>Off-path Caching</td>
<td>CINC, DCM</td>
</tr>
<tr>
<td>Popularity Caching</td>
<td>LCD, MCD, WAVE, CLS, MPC, DCM, MAGIC, PGBCS, LCLCS.</td>
</tr>
<tr>
<td>Non-Popularity Caching</td>
<td>LCE, Prob(p), Breadcrumbs, Probcache, Probcache+, CATT, Intra-AS, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, APP.</td>
</tr>
<tr>
<td>Distribution Caching</td>
<td>LCD, MCD, Probcache, Probcache+, WAVE, CATT, CLS, DCM, Intra-AS, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, LCLCS.</td>
</tr>
<tr>
<td>Non-Distribution Caching</td>
<td>LCE, Prob(p), Breadcrumbs, MPC, APP.</td>
</tr>
<tr>
<td>Probability Caching</td>
<td>Prob(p), Probcache, Probcache+, APP.</td>
</tr>
<tr>
<td>Non-Probability Caching</td>
<td>LCE, LCD, MCD, Breadcrumbs, WAVE, CATT, CLS, MPC, DCM, Intra-AS, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, LCLCS.</td>
</tr>
<tr>
<td>Redundancy Caching</td>
<td>LCD, MCD, Breadcrumbs, Intra-AS.</td>
</tr>
<tr>
<td>Non-Redundancy Caching</td>
<td>MCD, Prob(p), Probcache, Probcache+, WAVE, CATT, CLS, MPC, DCM, CBC, TLCC, MAGIC, CPHR, CINC, Edge, BidCache, PGBCS, APP, LCLCS.</td>
</tr>
<tr>
<td>Chunk Caching</td>
<td>Probcache, Probcache+, WAVE, CLS, MPC, Intra-AS, TLCC, CINC, PGBCS, APP.</td>
</tr>
<tr>
<td>Content Caching</td>
<td>LCE, LCD, MCD, Prob(p), Breadcrumbs, CATT, DCM, CBC, MAGIC, CPHR, Edge, BidCache, LCLCS.</td>
</tr>
</tbody>
</table>

6. CONCLUSION

In general, it is clear that caching strategies have a significant impact on the performance of the In-network cache, which manages the distribution of contents in different regions to make it easy to retrieve on demand.
In this review, the importance of the In-network cache has been clarified, in addition to clarifying the importance of the caching strategy that has a significant role in improving performance, such as reducing server load, congestion, and delay caused by users. Furthermore, caching strategies have been discussed and described clearly in terms of the principle and the purpose.

In this review, the categorizations of caching strategies are defined, then, the caching strategies are classified based on these categorizations that help the reader to understand the principle of caching strategy, and recognize the similarities and differences between these strategies.

REFERENCES


