DeepChunk: Deep Q-Learning for Chunk-based Caching in Wireless Data Processing Networks

Yimeng Wang, Yongbo Li, Tian Lan, and Vaneet Aggarwal

***Abstract*—A Data Processing Network (DPN) streams massive volumes of data collected and stored by the network to multiple processing units to compute desired results in a timely fashion. Due to ever-increasing traffic, distributed cache nodes can be deployed to store hot data and rapidly deliver them for consumption. However, prior work on caching policies has primarily focused on the potential gains in network performance, e.g., cache hit ratio and download latency, while neglecting the impact of cache on data processing and consumption.**

In this paper, we propose a novel framework, DeepChunk, which leverages deep Q-learning for chunk-based caching in wireless DPN. We show that cache policies must be optimized for both network performance during data delivery and processing efficiency during data consumption. Specifically, DeepChunk utilizes a model-free approach by jointly learning limited network, data streaming, and processing statistics at runtime and making cache update decisions under the guidance of deep Q-learning. It enables a joint optimization of multiple objectives including chunk hit ratio, processing stall time, and object download time while being self-adaptive under the time-varying workload and network conditions. We build a prototype implementation of DeepChunk with Ceph, a popular distributed object storage system. Based on real-world Wifi and 4G traces, our extensive experiments and evaluation demonstrate significant improvement, i.e., 52% increase in total reward and 68% decrease in processing stall time, over a number of baseline caching policies.

***Index Terms*—Data Streaming and Processing, Caching, Reinforcement Learning.**

I. Introduction

DATA collection, streaming, and processing are essential tasks for modern wireless networks due to the rapid development in areas such as Internet of Things, sensor networks, online data analytics and edge computing [1], [2], [3]. In such applications, massive volumes of data collected (and stored) by the network need to be streamed to multiple processing units to compute desired results in a timely fashion. One example of this type of application is intelligent transportation, where large volumes of sensor data and video footage are recorded from various monitoring points, and then fetched on-demand (through wireless networks) into distributed computing nodes, for applications ranging from vehicle identification to traffic analysis. Due to ever-increasing traffic in Data Processing Networks (DPN), [4], [5], [6], they often enhance the performance by caching data in nodes close to computing units and rapidly delivering those data for consumption.

Y. Wang, Y. Li, and T. Lan are with the School of Engineering and Applied Science, George Washington University, Washington, DC, 20052, USA, emails: {wangyimeng, lib, tlan}@gwu.edu. V. Aggarwal is with the School of Industrial Engineering and the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA, email: vaneet@purdue.edu.

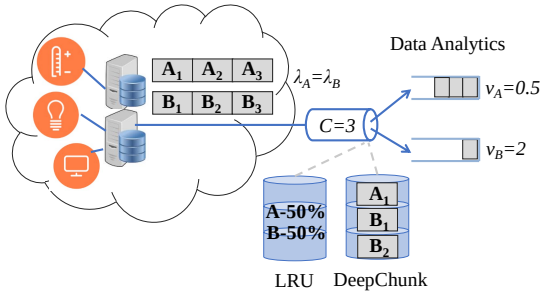


Fig. 1. An illustrative example of DeepChunk approach for data processing.

The design of caching policies, however, is primarily focused on the potential gains in network performance, e.g., cache hit ratio and download latency, while neglecting the impact on data processing and consumption. These policies include the Least Frequently Used (LFU) [7] and Most Popular Object [8] strategies that achieve a high cache hit ratio, and the Least Recently Used (LRU), qLRU, and kLRU [9] strategies that use request recency for cache update. Yet, for data processing networks, a data stream is simultaneously being fetched and consumed on the fly. Existing caching policies only considering network performance during data delivery fall short on addressing end-to-end tuple processing time [2], which depends on both data delivery and consumption. In practice, when available network bandwidth cannot fully sustain all data processing needs [1], [2], stalls during data processing become inevitable. As a result, the design of caching policies should be made aware of data consumption needs, to mitigate processing stalls and fulfill the timely online processing requirements.

In this paper, we propose a novel framework, DeepChunk, which leverages deep Q-learning [10] for chunk-based caching in DPN. Our key idea is that cache policies must be optimized for both (i) network performance during data delivery and (ii) processing efficiency during data consumption. Since the classical hit ratio metric does not account for partial files in the cache, we use the notion of *chunk hit ratio* that accounts for the existence of partial files in the cache. The necessity of taking data processing into account in caching can be illustrated via a simple example in Figure 1. Two data objects, A and B (each consisting of 3 chunks), are recorded on network edge and fetched by different data processing units through a wireless link that has a capacity of 1 chunk per second and is equipped with a cache of size $C = 3$. If A and B have the same request rate, they are equally likely to be stored in the cache

under LRU and LFU policies. Suppose that two applications with processing speed of $\nu_A = 1/2$ and $\nu_B = 2$ chunks per second start to request A and B at $t = 0$, respectively. It is easy to see that if A is cached, the processing of B stalls for a total of 2 seconds (i.e., 1 second waiting time for the first chunk and 1/2 second idle time before processing each subsequent chunks), and similarly the processing of A stalls for 1 second if B is cached. However, we show that stall-free data processing is indeed possible under an optimized, chunk-based caching policy. If the first 2 chunks of B are cached, we can fetch the last chunk of B by $t = 1$, so that the processing of B is stall-free. At the same time, caching the first chunk of A allows it to continue processing without interruption until the next two chunks are fetched by $t = 2$ and $t = 3$, respective. This caching policy not only achieves the same chunk hit ratio (i.e., 50%), but also minimizes end-to-end tuple processing time (i.e., 6 and 1.5 seconds for A and B respectively) with stall-free data processing. Thus, to fulfill the real-time data processing requirements, it is necessary to consider both network performance and data processing objectives.

A fundamental problem in our chunk-based cache system is the cache update policy, which entails two types of decisions at chunk level – how many chunks of a data object to admit (cache admission) and which chunks to evict from the cache (cache eviction) if it is already full – with the objective of jointly optimizing average chunk hit ratio, processing stall time, and object download time, in DPN. This optimization may be solvable if we can accurately model the correlation between the objective values and underlying variables, e.g., request arrival patterns, data popularity distributions, and wireless network conditions. However, this is very challenging and has not yet been well studied in the context of data processing networks, which represent a fairly complicated multi-point to multi-point system with the dynamics of data streaming and processing (from multiple data objects) closely coupled and jointly impacting the design objectives.

Hence, DeepChunk aims to develop an approach by jointly learning data streaming, processing, and network statistics and making decisions under the guidance of deep Q-learning [10]. We believe the approach is especially promising for chunk-based cache in data processing networks because: (i) it does not rely on precise and mathematically solvable models, which is hard to obtain in practical DPN, (ii) it is capable of supporting an enormously large state space, and (iii) it is self-adaptive to the dynamic environment, e.g., evolving data popularity/arrivals and time-varying wireless network conditions. In particular, the state space in DeepChunk's deep Q-learning includes data popularity distribution, cache states, request arrival statistics, network conditions and current request information, while its reward captures chunk hit ratio, processing stall time, and object download time. The output action determines DeepChunk's cache update policy, and the resulting reward is further fed-back to the neural network for learning.

We build a prototype implementation of DeepChunk with Ceph [11], a popular distributed object storage system. Ceph's cache node is modified to implement a deep Q-learning engine and a chunk-based cache module. Upon a request arrival, the cache module immediately streams all cached chunks and request the remaining chunks from the Ceph storage cluster, where all data objects are stored. The deep Q-learning engine obtains state updates $s(t)$, decides an action $a(t)$ based on the trained neural network, and then sends the action $a(t)$ to the cache module, which performs cache updates accordingly, calculate the resulting reward, and send it back to the deep Q-learning engine. To evaluate DeepChunk, we generate data popularities from Zipf distribution and utilize Linux TC traffic control [12] to emulate different wireless network conditions, based on real-world Wifi and 4G network traces [13]. We run extensive experiments to compare DeepChunk with a number of baselines including No-Cache, LRU, kLRU, gLRU. DeepChunk achieves up to 52% reward improvements compared with the baselines, and in all scenarios, it has the ability to balance different design objects, illuminating an interesting tradeoff between chunk hit ratio, processing stall time, and object download time.

The main contributions of this paper are as follows:

- We propose a novel framework, DeepChunk, which leverages deep Q-learning to optimize chunk-based caching in DPN. It only relies on obtaining limited network statistics on the fly to make cache update decisions and self-teach the optimal update policy.
- DeepChunk is able to operate in a fully unsupervised fashion with the objective of jointly optimizing both data processing and network performance, under the guidance of deep Q-learning.
- We implement a prototype of DeepChunk using Ceph, evaluate it on a real-world testbed, and compare its performance with a number of baseline caching policies. Significant improvement, i.e., 52% in total reward and 68% in processing stall time, is observed.

II. Related Work

Caching is widely used in networks for various objectives such as reducing latency, mitigating congestion, and improving user experience. Its applications include radio-access network [14], mobile 5G networks [15], web applications [16], BigData applications [17], distributed storage [18], [19], and video delivery [20], [21].

As a simple and intuitive cache replacement strategy, LRU-based caching mechanisms have been widely studied [22], [23]. One of the key issue in LRU-based caching strategies is that an arrival of a large file can evict multiple small files [24]. Several different approaches have been proposed to improve its performance performance with realistic, vaiable file sizes and popularities. The qLRU algorithm [25] stores the newly arrived object and evicts the LRU object with probability q, and the kLRU algorithm [25] only executes the storing/evicting procedure when a new object has been advanced in all $k - 1$ virtual LRU caches ahead of the "physical" cache. Other than the LRU-based algorithms, the R-UPP and P-UPP algorithms [26] specifically consider user preference profiles, while AdaptSize algorithm in [24] is optimized for the case where the object sizes are widely

varied. In addition, the Least Hit Density (LHD) policy [27] predicts each object's expected hits-per-space-consumed to improve cache replacement strategy. In contrast, in this paper, we propose a chunk-based caching framework that leverages deep Q-learning [10] and optimizes replacement strategy at chunk level. The use of deep Q-learning allows us to develop a model-free approach and efficiently compute the optimal chunk replacement strategy to maximize multiple objectives in a unified framework. Deep Q-learning has been successfully applied to address network optimization problems in many different areas such astraffic engineering [28], video bitrate control [29], LTE femtocell configuration [30], ride-sharing [31], and cell outage management [32].

III. Background and Problem Statement

We consider a DPN as a set of data sources and processing units, connected through a wireless network to apply various data operations and computations. Massive volumes of data collected (and stored) by the network often need to be streamed to multiple processing units and processed in real time. This feature is also known as stream data processing [2] and can be found in many existing and emerging applications such as the Internet of Things, sensor networks, online data analytics and edge computing [1], [2]. In such DPN, cache nodes can be deployed between the data source and the processing units, to store and reuse hot data objects passing through the network. Most of the existing cache replacement policies, such as LRU and LFU [33], focus on network performance metrics such as cache hit ratio. However, in DPN, as demonstrated in the previous example in Figure 1, higher hit ratio does not necessarily lead to more efficient data processing, as measured by processing stall time. Motivated by this phenomenon, we acknowledge the new cache design challenges arising from DPN and develop a chunk-based caching framework guided by deep Q-learning.

To reduce data processing stall, caching the starting chunks of different data objects is crucial, while storing the complete data objects does not offer any additional benefits, as illustrated by the example in Figure 1. It mandates us to consider chunk-based caching in DeepChunk. At the core of chunk-based caching policies is the need to make cache admission and eviction decisions for each individual data chunk, resulting in higher decision complexity on the fly. In object-based policies, such as kLRU and qLRU [34], cache admission and eviction decisions are typically binary, as to whether or not to add a new object to replace the least-recently-used one in the cache. A chunk-based caching policy, however, must decide *how many* new chunks to admit into the cache when a requested data object traverses the cache node, even-though the cache eviction can employ a least-recently-used strategy.

Consider the example in Figure 1. Suppose that the cache currently contains 2 chunks of A and 1 chunk B, i.e., A_1, A_2, B_1. If a new request of A arrives with a data processing speed of $\nu_A = 0.5$, stall-free processing can already be achieved with the 2 chunks of A in the cache (ignoring a fixed initial waiting time to start streaming). There is no need to admit any more chunks of A into the cache, as it only negatively impacts the processing stall time of other objects that must be evicted as a result. On the other hand, if a new request of B arrives with a data processing speed of $\nu_B = 2$, the processing will stall, demanding new chunks of B to be admitted into the cache. However, we need to carefully choose the number of object-B chunks to admit, since an overly aggressive policy (e.g., adding both B_2 and B_3) would superfluously stall future processing requests of A. In DeepChunk, we collect runtime statistics from the DPN – including cache state, request rates, data processing speeds, chunk sizes, and network configuration – and leverage deep Q-learning to guide the chunk-based decision making.

IV. Proposed DeepChunk Framework

DPNs focus on rapid processing of incoming data streams. Hence, traditional design policies, maximixing hit ratio and hit rate, often do not necessarily lead to more efficient data processing, as demonstrated in our illustrative example, in Section I. Motivated by these observations, we propose a new framework to optimize DPNs for improved Quality of Experience (QoE). In this section, we introduce our chunk-based caching policy and then optimize it for DPNs.

We assume that there are N data objects to be streamed and processed by different processing units. Each object $i \in \{1, 2, ..., N\}$ is partitioned into F_i identical-sized chunks. A cache node is located in close proximity to the processing units and can store up to C data chunks. We consider a time-slotted system model, in which each time bin contains exactly one request arrival. At a given time bin t, let $C_i(t)$ denote the number of chunks of data object i stored in the cache node, satisfying $0 \leq C_i(t) \leq F_i$. The cache size constraint requires $\sum_{i=1}^{N} C_i(t) \leq C$ for any time bin t. For the proposed caching strategy, in steady state the above cache size constraint will hold with equality since we will not waste any cache capacity in the steady state.

Our chunk-based caching policy is formulated as follows. In time bin t, when data object i is requested, $C_i(t)$ chunks stored in the cache are directly streamed to the processing unit, which then immediately begins data processing. At the same time, the remaining $F_i - C_i(t)$ requested chunks will be delivered from the data source and through the network. When the complete data object i is not yet in the cache, i.e., $C_i(t) < F_i$, E_c additional chunks of object i satisfying $0 \leq E_c \leq F_i - C_i(t)$ will be added to the cache. Thus, the number of object-i chunks in cache increases from $C_i(t)$ to $C_i(t+1) = C_i(t) + E_c$ in the next time bin $t+1$. Further, these $C_i(t+1)$ chunks of object i are moved to the head of line in the cache since they become most-recently-used. It is easy to see that to mitigate processing stall, we should always place the first $C_i(t+1)$ chunks of object i in cache and stream them to jump-start data processing. Finally, when the cache is already full, to make space for the added chunks, an equivalent number of chunks must be removed from the cache. In DeepChunk, we adopt a policy similar to LRU and remove a necessary number of chunks from the tail of the cache line.

Our goal is to design a cache policy to optimize both network performance and data processing objectives. Let $h(t)$ be

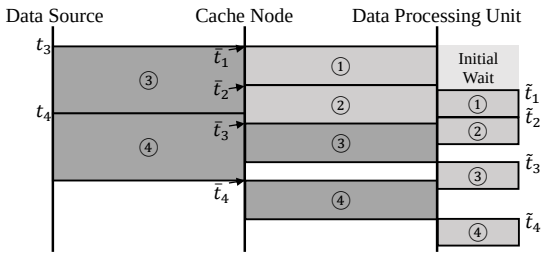


Fig. 2. Calculation of processing stall time.

the chunk hit ratio in time bin t (defined formally later in this section), $T_d(t)$ the average object download time, and $T_s(t)$ the average processing stall time. Specifically, DeepChunk aims to optimize the following objective in steady state (for sufficiently large τ) over feasible cache policies $\mathcal{P}$:

$$\max_{\mathcal{P}} \frac{1}{\tau}\sum_{t=1}^{\tau} \alpha_1 h(t) - \alpha_2 T_s(t) - \alpha_3 T_d(t), \tag{1}$$

where $\alpha_1, \alpha_2, \alpha_3$ are non-negative weights assigned to chunk hit ratio, processing stall time, and object download time, respectively. In practice, we can adjust these weights to achieve different tradeoffs between the network performance and data processing objectives. For instance, when $\alpha_1 = \alpha_3 = 0$, the resulting cache policy minimizes processing stall time $T_s(t)$. We use Reinforcement Learning (RL) to solve the above optimization, and for each network state, to determine the optimal cache replacement, i.e., the number of chunks E_c to replace. We note that existing caching policies typically rely on constant cache replacement strategies, e.g., $E_c = F_i$ for the LRU policy and $E_c = 1$ for the gLRU policy [23], thus lacking the ability to adapt cache replacement on the fly with respect to dynamics in DPN.

Our goal in this paper is to jointly optimize both the network performance and the data processing efficiency in a DPN. To tackle this multi-objective optimization problem, we introduce three metrics that will be computed from limited runtime statistics on the fly, and then leverages RL to develop an automated solution, DeepChunk.

Process stall time. To find processing stall time, we assume that the data processing unit is equipped with a sufficiently large buffer, so all streamed data chunks are consumed by the processing unit without the need for retransmission. For the simplicity of notations, we drop the index t in the following derivations, while the variables such as cache state C_i and wireless bandwidth r_1 are indeed time-varying.

Consider a request in time bin t. Since the first C_i out of F_i chunks of data object i are already stored in the cache and the remaining $F_i - C_i$ chunks need to be streamed from the data source, we find the processing stall time by considering a two-stage buffer problem. As shown in Figure 2, let r_2 denote the available bandwidth (i.e., data streaming speed) from data source to cache node, r_1 denotes the speed from cache node to processing unit, and ν_i denotes the data processing/consumption speed of object i. Note that for wireless channels with time-varying bandwidth, $r_1(k)$ denotes the average speed when transmitting the kth chunk. We consider 2 stages in data streaming, t_k and $\bar{t}_k$ to denote the time when the kth chunk starts streaming from data source to cache node and from cache node to processing unit, respectively. Since the first C_i chunks are already stored in the cache and the other chunks are streamed one-by-one from data source, we have

$$t_k = \begin{cases} t_{k-1} + \frac{1}{r_2}, & k \in [C_i+1, K], \\ 0, & k \in [1, C_i]. \end{cases} \tag{2}$$

Next, each data chunk k can be streamed from the cache node to the processing unit, when it becomes available (i.e., at $t = 0$ for any cached chunk and $t_k + 1/r_2$ if it needs to be fetched from data source) and after the preceding chunk $k-1$ is delivered (i.e., at $\bar{t}_{k-1} + 1/r_1$). Combining these, we have

$$\bar{t}_k = \begin{cases} \max\{\bar{t}_{k-1} + \frac{1}{r_1(k-1)}, t_k + \frac{1}{r_2}\}, & k \in [C_i+1, F_i], \\ \bar{t}_{k-1} + \frac{1}{r_1(k-1)}, & k \in [2, C_i], \end{cases} \tag{3}$$

except that $\bar{t}_1$ of the first chunk depends on whether any chunks of object i are cached, i.e.,

$$\bar{t}_1 = \begin{cases} t_1 + \frac{1}{r_2}, & C_i = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

Finally, the processing unit consumes data at speed ν_i. Then the processing start time of chunk k can be recursively computed from $\bar{t}_k$'s as

$$\tilde{t}_k = max\{\tilde{t}_{k-1} + \frac{1}{\nu_i}, \bar{t}_k + \frac{1}{r_1(k)}\}, \text{ and } \tilde{t}_1 = t_{ini}, \tag{5}$$

where t_{ini} is the initial wait time (or startup delay) mentioned in Section III. When $\tilde{t}_k > (k-1)/\nu_i + t_{ini}$, the processing unit would experience stall time waiting for chunk k. The total processing stall time of object i is then given by the difference between actual play time and expected play time of the last chunk F_i (since stall time accumulates during processing), that is

$$T_s = (\tilde{t}_{F_i} - \frac{F_i - 1}{\nu_i} - t_{ini})^+. \tag{6}$$

Object download time. The download time of each data chunk k can be found through $\bar{t}_k$, which is the time to starting streaming chunk k from the cache node to the processing unit and is already given in the above analysis of processing stall time. Thus the download time T_d of data object i that is equivalent to the arrival time of last chunk F_i can be derived directly from $\bar{t}_{F_i}$ as follows:

$$T_d = \bar{t}_{F_i} + \frac{1}{r_1(F_i)}. \tag{7}$$

Chunk hit ratio. We note that for object-based cache systems, a "hit" occurs if the requested data object is found in the cache. Thus, the hit ratio equals to the probability that the requested data object is stored in the cache. For chunk-based caching considered in this paper, we define a similar *chunk hit ratio* to quantify the performance of the proposed caching policy. Specifically, we denote $h_i = C_i/F_i$ as the chunk hit ratio of data object i, which is the percentage of object-i chunks that are stored in the cache and can be directly used to serve a request. This notion of chunk hit ratio generalizes

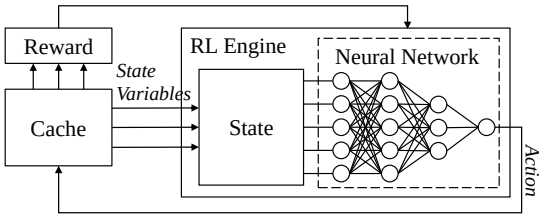


Fig. 3. Reinforcement learning mechanism of DeepChunk decision making.

the existing object-based definition, which is binary – $h_i = 0$ for a cache miss and $h_i = 1$ for a cache hit – and cannot be used to evaluate chunk-based caching systems. Finally, the overall chunk hit ratio is simply the average of chunk hit ratio of individual data objects weighted by request arrival rates λ_i, i.e., $(\sum_i \lambda_i h_i)/(\sum_i \lambda_i)$.

V. Reinforcement Learning

A. Deep Q-Learning

To make caching decisions in different system states, we utilize deep Q-learning to dynamically generate optimized values.

With the development of the neural networks, deep Q-learning is commonly used in modern decision making tasks. The advantages of RL decision making process are: (i) supporting enormously large state space, (ii) scalable to different input dimensions, and (iii) self-adapting to the environment including evolving data popularity and varying network conditions. We take advantage of RL to adapt to these dynamic features in the caching problem.

Figure 3 illustrates how deep Q-learning is solving our cache decision problem. At each time bin t, the cache node monitors the current state $s(t)$ of the system. When a request $q(t)$ is observed at the cache node, the RL Engine feeds the current state $s(t)$ into a neural network to generate an action $a(t)$. Further, according to the state $s(t)$ and the action $a(t)$, the state will be pushed to the next state $s(t+1)$. When the next request arrives, the reward $r(t)$ of the previous action $a(t)$ can be observed, further fed back to train the neural network.

To dynamically improve the reward, Q-values for each state-action pair are maintained to represent the potential reward of decisions. The Q-values are updated by the immediate reward, and an expected optimal future reward which is discounted by a factor $\gamma, (0 \leq \gamma \leq 1)$:

$$Q(s(t), a(t)) = r(t) + \gamma \max_a Q(s(t+1), a). \tag{8}$$

To maintain a large system state space, an artificial neural network is utilized. At time t, the detected state $s(t)$ are sent into the neural network as the input. After the process in hidden layers, the Q-values for all possible actions are observed at the output layer. The decision making policy follows the *Epsilon Greedy* scheme: with probability $1 - \epsilon$, the agent will choose the action that results in the highest Q-value, otherwise, select a random action with equal probability. The ϵ reduces linearly from 1 to 0.1 over iterations, thus the agent starts with an aggressive exploring behavior, then tends to utilize the learned experience over time.

After the action is selected, according to the corresponding reward $r(t)$, the Q-value is updated with a learning rate β:

$$\begin{aligned} Q'(s(t), a(t)) \leftarrow (1-\beta) Q(s(t), a(t)) \\ + \beta[r(t) + \gamma \max_a Q(s(t+1), a)]. \end{aligned} \tag{9}$$

Similar to ϵ, the learning rate β is also reduced linearly.

The neural network contains multiple layers to estimate the Q-values of each state-action pair. For different problems, specific types of neural networks – such as Convolutional Neural Network (CNN) [35], Recurrent Neural Network (RNN) [36], etc. – can be used to improve the training speed and the estimation accuracy. For our proposed model, we choose a simple neural network which contains only fully connected layers as the hidden layer.

B. Reward, States, and Actions

To use deep Q-learning for decision making, we define the three crucial elements - **state**, **action** and **reward** - as follows:

1) Reward: We consider both data processing and network performance to be optimized as the reward. As mentioned in Section IV, *chunk hit ratio*, *processing stall time* and *object download time* are defined as the reward variables. We apply three weight factors α_1, α_2 and α_3 to adjust the importances of the three reward variables. Thus, the immediate reward is defined as:

$$r = \alpha_1 h - \alpha_2 T_s - \alpha_3 T_d. \tag{10}$$

Note that when the algorithm is running, all three objectives – h, T_s, T_d – can be measured from the cache node or the data processing units, while Equations (6) and (7) can be utilized to pre-train the neural network from zero knowledge, to improve the speed of convergence.

2) States: The state variables should reflect the system status, further affect the reward feedback of different actions. We measure the system state in our caching model as a five-tuple: $(\overrightarrow{p}, \overrightarrow{n}, \overrightarrow{c}, \overrightarrow{o}, f)$.

Specifically, f denotes the *currently requested object.* As we designed, only this object will be cached in the same time iteration. $\overrightarrow{p}$ denotes the *popularity distribution* over all objects. In practical networks, the popularity of objects are changing over time. In order to adapt to the latest popularities, instead of using a static distribution, we maintain a sliding window to monitor popularities of all objects during the past n requests:

$$w_p = [f(t-n+1), f(t-n+2), ..., f(t)]. \tag{11}$$

Similarly, the *network condition* history $\overrightarrow{n}$ is computed through another sliding window to track the time-varying bandwidth of wireless channel:

$$w_n = [b(t-n+1), b(t-n+2), ..., b(t)], \tag{12}$$

where $b(t)$ denotes the average bandwidth in time slot t. The use of sliding windows can eliminate (i.e., average out) system randomness at small timescale, while allowing us to track the trend/change in object popularity and network bandwidth at

large timescale, providing important information for the cache optimization. The sliding windows contain the histories of the past requested objects/network bandwidths for n time slots. The currently *cached chunks* for all objects are denoted by $\overrightarrow{c}$. These state variables are utilized to determine how data chunks in the cache are updated, in order to optimize the reward. Finally, the *order* of requested objects is expressed by an array $\overrightarrow{o}$. In our DeepChunk policy, we make decisions to replace the least-recently-used object chunks by new chunks. This state variable will indicate chunks from which object/objects will be removed when making decisions.

All the elements in this four-tuple can be obtained at the cache node when a request arrives. For cache decision, the four-tuple is pushed into the input layer of the neural network.

3) Actions: Current RL applications can have a large state space, but it is not well scalable with action. So, we define an unsophisticated action to fit our problem to this feature. The action $a(t)$ represents *the number of chunks of the requested object f to be added to the cache*. When the action is made, E_c additional chunks of object f will be stored in the cache storage. When the cache storage is full, chunks from the LRU objects will be removed.

At a certain state $s(t)$, the action $a(t)$ will result in a cache increment of the requested object f, and an up to $a(t)$ cache decrement of the least-recently-used objects. Thus, in the next time slot $t+1$, the cached chunks $\overrightarrow{c}$ depends on the cache decision $a(t)$. Other state variables – the popularity distribution $\overrightarrow{p}$, network condition $\overrightarrow{n}$, request order $\overrightarrow{o}$, and requested object f – will follow their own randomnesses.

C. Algorithm Training

We train the deep Q-learning algorithm using the simulated rewards (from Equations 6, 7, and 10). Following the Zipf distribution, an object request is randomly generated in each time iteration. Depicted in Figure 3, the five-tuple $(\overrightarrow{p}, \overrightarrow{n}, \overrightarrow{c}, \overrightarrow{o}, f)$ is observed by the state listener in the RL Engine, and further fed into the input layer of the neural network. After the calculation of the neural network, the estimated Q-values of all possible actions are obtained at the output layer. The action with the highest Q-value, or a random action will be chosen according to the Epsilon Greedy policy. Finally, using the calculated Q-value shown in Equation 9, the neural network is updated.

Reinforcement learning has been developing in order to address varied problems. To adopt to different application scenarios, advanced Q-learning algorithms are utilized, including Dueling-DQN [37], Double-DQN [38], Deep Recurrent Q-learning [39], Rainbow DQN [40], etc. We test our proposed cache environment in both the original deep Q-learning and Dueling-DQN – which separately estimates state values ($V(s)$) and advantages ($A(s,a)$) – algorithms. The training curves are shown in Figure 4. According to the figure, Dueling-DQN does not improve the final reward of the training process comparing with the original deep Q-learning algorithm since every action making in the service placement problem impacts the instant reward critically in almost all kinds of states. However, because the training speed suffers from extra layers utilized to separately estimate both $V(s)$ and $A(s,a)$, the convergence of DDQN is slower than deep Q-learning. Thus, to accelerate the training speed, we choose the original deep Q-learning method for the proposed optimization task.

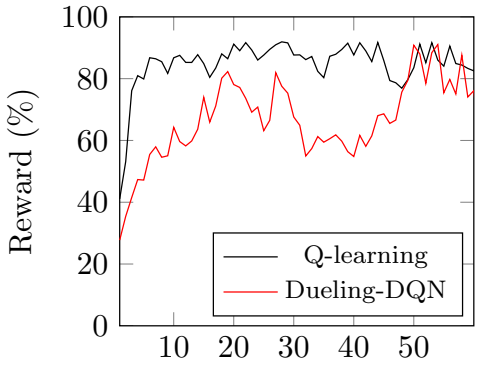


Fig. 4. Training curves of deep Q-learning and Dueling-DQN algorithms.

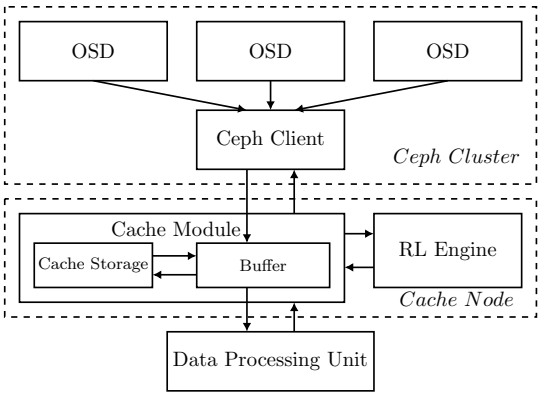


Fig. 5. Diagram of implemented deep Q-learning prototype.

VI. Implementation

In this section, we will describe the prototype implementation details. In specific, we will discuss the DPN setup in terms of the following aspects: data source, cache node, data processing unit, and the links between nodes. The system diagram is depicted in Figure 5.

A. Data Source

Three virtual machines running a Ceph [11] cluster are implemented as the data source. All files used in the experiments are divided into chunks. The files are stored in the three Object Storage Daemons (OSDs). Another virtual machine within the same cluster configuration is set as a Ceph client, which is responsible to collect chunks (Ceph objects) from the storage cluster and send them to the cache node. A Tornado server [41] is implemented on the Ceph client, where it handles the fetch requests from the cache node with responses from the data source. The transmission latency and processing delay within the Ceph cluster machines (those for cloud storage, Ceph client, and Tornado server) are ignored.

B. Cache Node

The cache node aims to simulate an edge node between the data processing unit and the source. When the processing unit sends a request to the cache node, the cache node fetches the chunks of the object that do not exist in the cache from the Tornado server at the source, and delivers the chunks that obtained (previously cached and newly downloaded) to the processing unit simultaneously. Further, with the arriving chunks from the source, the cache node runs an RL Engine to determine the updated placement of the chunks of different objects in the cache. Cache node has two modules - Cache Module and Reinforcement Learning Engine, which together achieve the functions explained above. In our experiments, one virtual machine is utilized to run both the modules.

1) Cache Module: The Cache Module manages the placement of the chunks on the cache node, fetching the chunks from the data source, and delivery of contents to the processing unit. A Tornado server is used to build the connection with the data processing unit, and a Tornado client is paired with the server on the data source. When an object request is received from the processing unit, the cache module starts to push the cached chunks and fetch the missing chunks from the source simultaneously.

A buffer is used to store the fetched data bytes from the data source. The fetched contents will stay in buffer till all the chunks are sent to the processing unit. The cache policy will determine how many of these contents will be transferred to the cache storage. After the transmission of the contents to the processing unit and the transfer of required contents to the cache storage, the contents are removed from the buffer.

The metric of chunk hit ratio is collected from the cache module. We will next describe the Reinforcement Learning Engine that makes the decision for the update of caching policy which will influence how many of the chunks fetched from the data source will go to cache storage and which contents will be removed from the cache storage.

2) Reinforcement Learning Engine: The Reinforcement Learning (RL Engine) implements the cache update policy. To accelerate the training process, the neural network is pre-trained using simulated inputs. A series of Zipf-random requests are used to activate the evolution of the system state. After the action is made, the reward is calculated by the monitored chunk hit ratio and calculated stall/download time from Equations (6) and (7). The pre-trained neural network is then stored in the engine. When a cache decision is needed, the Cache Module will consult the RL Engine. In our implementation, the RL Engine and Cache Module are located on the same node.

A state listener queries a message from the Cache Module which contains all the required state information, including current cache storage status, request history, network bandwidth history, and currently requested object. By feeding the state variables to the neural network, the RL Engine obtains the action. The action is then sent to the Cache Module via a message to transmit the cache update decision.

For the real-time training of RL Engine, the reward variables (the chunk hit ratio from the cache module, the download and stall time from the data processing unit) are collected after all chunks are received by the processing unit. We note that the obtained reward is dependent on the previous states and actions. In the evaluations, it takes 60 minutes to train the neural network with a million preset samples. The training process is controlled by a linear control signal, which decreases the epsilon greedy parameter ϵ and the learning rate β linearly.

C. Data Processing Unit

A Tornado client is set up as the data processing unit. The processing unit continuously sends object requests to the cache node following a Zipf distribution. A Zipf "seed" is utilized to order the popularities of objects. Each request is sent to the cache node, and the object is received from the cache module. The data processing unit records the reward attributes, including processing time, initial waiting time, and stall time. These reward variables are sent to the cache node, which will be used by the RL engine for online learning and improving of the caching policy.

D. Data/Signal Flow

In the above, we introduced the function modules. Now, we will show an example to trace the data flow, and further describe how the system works.

When initiated, the Tornado servers in the Ceph client and the cache module, and the state listener in the RL Engine are activated. The data processing unit generates a random integer following Zipf distribution and decides which object to be requested according to the Zipf seed. Then a request of the object is sent to the cache module.

The cache module maintains awareness of cache status in real time. Once the new request is detected, it is able to build up the system state by cache storage status (cached chunks of all objects), updated request history, and the currently requested object. We denote this state as $s(t)$ for clearer demonstration. $s(t)$ is further pushed to the RL Engine via a message. Then, the action listener is activated.

The RL Engine captures state message by the state listener and makes an action $a(t)$ based on previously trained neural network. The RL Engine then sends $a(t)$ to the cache module through a message and starts to listen for the reward.

The cache module obtains the action message $a(t)$, and begins the data flushing process. It first reads all cached chunks of the requested object to its buffer, then sends a request for the missing chunks from its Tornado client to the server in Ceph cluster.

At the Ceph cluster, the chunks stored in Ceph OSDs are fetched as the response data. When sending the response, the Tornado server will flush the chunks into the network interface one by one to achieve a streaming feature.

The cache module starts two threads simultaneously. The fetching thread pushes the fetched bytes into the buffer. It is also responsible to write a part of the chunks into files according to $a(t)$, and further store them into the cache storage. The flushing thread keeps flushing the existing bytes in the buffer to the network interface. When the buffer is empty, it waits for the fetching thread until another chunk is ready in the

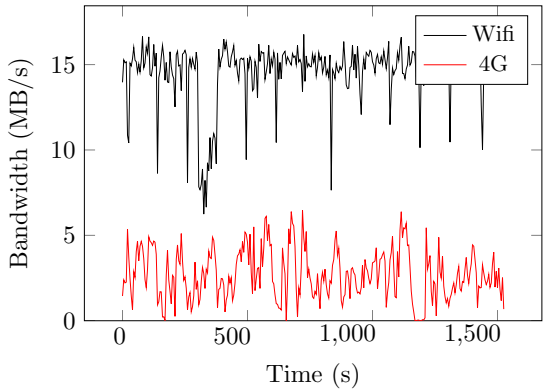


Fig. 6. Bandwidth traces of Wifi and 4G networks.

buffer. After all chunks are fetched and flushed, the connection finishes. Then, the reward listener enters standby mode.

The data processing unit receives the data stream. A timer will measure the download time from the moment when the request is sent, to the moment when the connection is finished. By subtracting the preset initial wait time and processing time, it obtains the processing stall time. A message carrying stall time and download time is reported to the cache module.

The reward value is calculated at the cache module. Note that three terms: chunk hit ratio, processing stall time, and object download time. The chunk hit ratio is obtained from the cache status which can be found in $s(t)$. Further, the reward $r(t-1)$ is sent to the RL Engine.

At the RL Engine, $r(t)$ is utilized to train the neural network. In our deep Q-Learning algorithm, the reward $r(t-1)$ is correlated with the previous state and action, $s(t-1)$ and $a(t-1)$. After the whole process is complete, the data processing unit loops.

VII. Evaluation

In this section, we present our evaluation results based on the system implementation described in the last section. We will leverage real-word Wifi and 4G traces [13] to compare DeepChunk with a number of baselines, including LRU, k-LRU, gLRU, and No-Cache policies.

A. Configuration

1) Machine Setup: Virtual machines running the Ceph cluster have identical disk space which is $10GB$. The virtual machine acts as the cache node has $256GB$ disk, $16GB$ memory, and has a core of Intel Xeon CPU E5-2630 v3 (2.40Ghz).

2) Link Setup: We use two wireless bandwidth traces (for Wifi and 4G channels, respectively) in our experiment. For the Wifi trace, we download a 100GB file on a computer, and measure the real-time downloading rate. For the 4G channel, we use the database collected by Ghent University [13]. The bandwidth traces are depicted in Figure 6, with an average bandwidth of 15MB/s for the Wifi channel and 2.6MB/s for the 4G channel.

All experiment machines are within the same local network. We use the Linux TC traffic control feature [12] to throttle the bandwidths between nodes. The wired link between the cache node and the cloud server node is set to 1.5MB/s, and the wireless link between the user client and the cache node is dynamic following the Wifi/4G bandwidth traces with average rates of 15MB/s and 2.6MB/s. The bandwidths between Ceph nodes (OSDs and Ceph client) are not a bottleneck, so no additional bandwidth restriction is imposed.

3) Objects: We use video files to run the experiments. Each file consists of multiple chunks, where each chunk is 1MB. Since the file sizes are not multiples of 1024^2, the last chunk of each file can be less than 1MB. Further, the processing speed of the files is divided into two groups – 1 MB/s and 3 MB/s – to represent different data processing applications.

We generate file popularities using Zipf's distribution [42] with the parameter $z = 1$. We sort the files by their sizes as the popularity ranks. In general, the file sizes and the popularities are dependent on each other. We consider a positive correlation between the file size and the popularity, in which larger file sizes have higher popularity. The popularity numbers are taken as the arrival rates of the requests at the user client.

4) Learning Parameters: A neural network is built by three parts: the input layer, the hidden layers, and the output layer. For the proposed optimization problem, we utilize four fully connected layers as the hidden layers. Each fully connected layer contains 1024 neurons.

Described in Section V-A, the learning rate β and the decision parameter ϵ is reduced linearly. Among the learning process, the learning rate β is reduced from 10^{-5} to 10^{-7}, and the decision parameter ϵ from 1 to 0.1.

5) Evaluated Policies: We compare the proposed DeepChunk policy with four baseline strategies, as described below.

No-Cache: This caching policy does not store anything in the cache.

LRU: The LRU caching policy [43] moves the requested file to the head of the cache, if already in the cache. If it is not in the cache, the file is added to the head of the cache and the files are removed from the tail to make space for the incoming file. Due to different file sizes, multiple files can be evicted to make space of a large incoming file.

kLRU: Instead of caching every missed request, kLRU [34], [9] deploys k virtual LRU caches ahead of the physical LRU cache to achieve a selective caching scheme. The virtual caches cache only file pointers instead of the data. A virtual/physical cache will store the pointer/data only if there is a hit in the LRU cache ahead of it, and replace the object following the LRU policy. In our experiments, two kLRU policies are tested. kLRU-1 and kLRU-2 policies apply 1 and 2 virtual caches ahead of their physical cache, respectively.

gLRU: Distinct from the previous policies, the generalized LRU [23] extends the LRU caching algorithm from file-level to a chunk-level algorithm. Upon a request arrival, if the requested file is not completely in the cache, one additional chunk of that file will be stored. When the cache storage is full, one chunk of the LRU object will be replaced. Further, all chunks of the requested file will be moved to the head of the cache (in order, such that the earlier chunks are towards the head so that they are evicted later).

TABLE I
REWARD SUMMARY FOR POLICIES IN DIFFERENT POPULARITY AND CACHE SIZE SETTINGS.

Policy	Reward	
	Wifi connection,	4G connection,
DeepChunk	-11.7070387055	-50.0106189471
gLRU	-17.3062326451	-68.7043762353
kLRU-1	-24.4986792735	-63.9718621500
kLRU-2	-22.8175804253	-60.8568648048
LRU	-20.5943264981	-70.3070151289
No cache	-31.6773507333	-93.8548818358

B. Evaluation Results

In this subsection, we compare the proposed DeepChunk policy with the baseline policies stated above. All strategies are run on both the Wifi and 4G environments. The cache size is set to 80MB (80 chunks), reward factors are set to 10 for the chunk hit ratio, 1 for the processing stall time, and 0.5 for the object download time. Under this reward setting, increasing the hit ratio of an object by 10%, decreasing the processing stall time by 1 second, and decreasing the total download time by 2 seconds are considered the same amount of contribution.

The experiment rewards of tested policies are shown in Table I. DeepChunk policy improves the total reward by 32.35% to 52.20% as compared to the other caching algorithms on the Wifi channel, and by 17.83% to 28.87% on the 4G channel. The other chunk-level policy, gLRU, suffers a lower reward compared with the file-level kLRU policies on the 4G channel. This fact indicates that without a proper joint optimization that takes into account time-varying wireless channel conditions, chunk-level cache policies alone cannot achieve optimal performance, as evidenced by the superior performance of our proposed DeepChunk policy.

Figure 7 shows the reward breakdown for the Wifi connection. From the figure, we observe that our DeepChunk policy outperforms all other policies in terms of all three reward factors. Compared with the file-level policies, the improvements are up to 44.44% on hit ratio, 68.10% on stall time, and 15.31% on object download time. Our policy did not improve significantly on download time since it's considered a minor contributor to the total reward. For the stall time and the hit ratio, the improvements are more obvious.

Compared with the other chunk-level policy gLRU, our policy is also better for all three reward factors. We observe that the gLRU policy also promises lower stall time. This result shows that by bringing the cache strategy into the chunk level, stall time, a major contributor to user experience, can be improved. Although gLRU has the lowest hit ratio among all the strategies, it's total reward is still better than all other file-level policies. However, without the learning feature to decide how many chunks should be cached, it cannot reach the optimal.

Similar results are concluded from Figure 8 for DeepChunk. When the connection is on 4G, DeepChunk improves the reward factors by up to 45.71% on hit ratio, 29.64% on stall time, and 19.43% on download time compared with file-level policies. However, in this experiment, the performance of gLRU policy is not ideal. Its hit ratio is still the lowest (changing bandwidth does not affect the hit ratios of caching policies), and the stall time becomes longer than the kLRU policies. This is because when the bandwidth is low, more chunks need to be stored to eliminate the stall time. The gLRU policy caches 1 chunk for each request, resulting in a very slow growth in the number of cached chunks.

In Figure 9, we show the improvement obtained by increasing the cache size. Take the LRU policy as an example, both stall time and chunk hit ratio are improved by 50% when the cache size is increased from 40 chunks to 80 chunks. With the increasing cache size, the average stall time improves smoothly, while the chunk hit ratio does not. It is because a slight increase in cache size is not able to make the storage capacity for one additional big file.

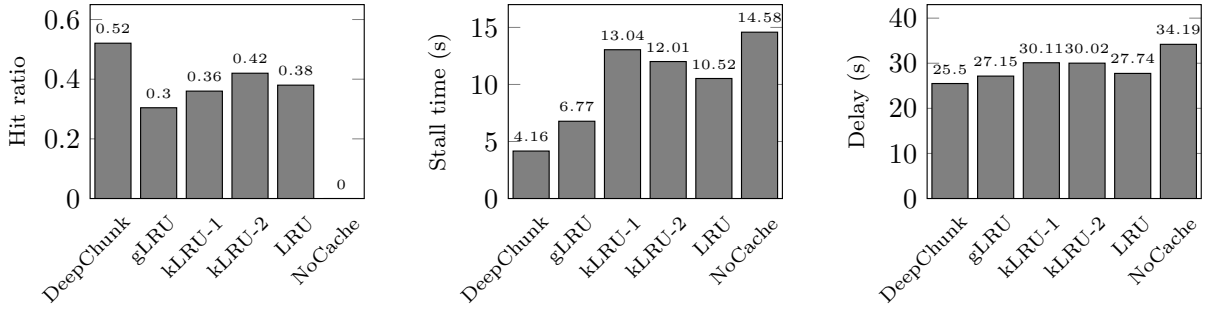
Figure 10 describes the importance of cache chunk decision optimization. We run the DeepChunk policy at the cache node with an empty cache storage, while a file with 40 chunks is repeatedly requested. As the time evolves, the chunk hit ratio of individual requests will grow from 0 to 1 linearly. The x-axis in the figure is marked by the chunk hit ratio of the request. As it grows, the download time decreases linearly. However, with the chunk hit ratio at 0.4 (iteration 10), the stall time hits 0 which is its minimum. Thus at this state, caching more chunks of this file will no longer gain rewards from the stall time, which has a heavy weight. The reward curve shows the same information, the growth slows down after the 10th iteration. Since DeepChunk is a state-aware policy, it tends to cache fewer chunks to save cache space for other files when the stall time can no longer be improved.

We extract the reward variables for individual files while running the experiments. In Figure 11, we compare the average stall time and chunk hit ratio for each file when different caching policies are applied. According to Figure 11 (a), the five files experience similar stall time under the DeepChunk policy. The standard deviation is 0.268, that is lower than LRU's standard deviation which is 2.189. As depicted in Figure 11 (b), although chunk hit ratio is not the largest weighted term in the reward function, DeepChunk still has a lower standard deviation (0.074) as compared to LRU (0.104). We can conclude that under the DeepChunk policy, more space is saved to reduce the overall stall time among all files.

Finally, we measure the performance of the policies when the popularity distribution suddenly changes. At the user client node, we change the popularity ranking for each of the 50 requests. From Figure 12, we observe that since the DeepCache keeps monitoring the historical probability distribution of the requests in a sliding window, and its performance (both processing stall time and chunk hit ratio) is better than the static LRU algorithm. This verifies that our DeepChunk is more robust and has the ability to adapt to time-varying data popularity in a dynamic environment.

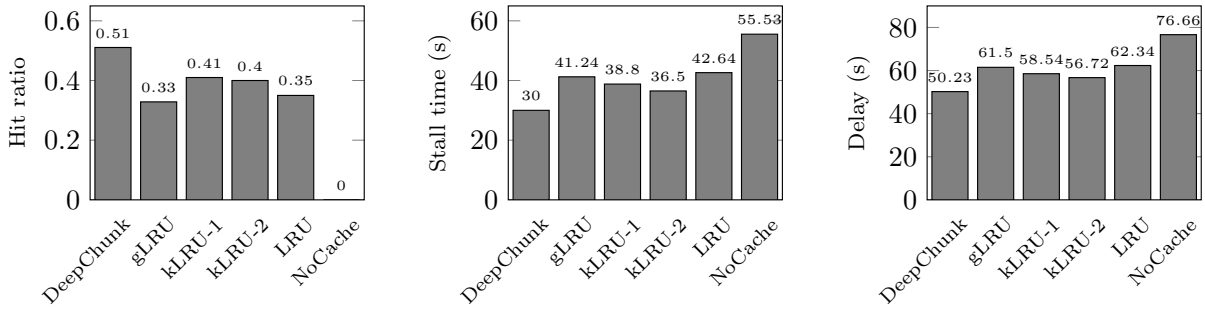
VIII. CONCLUSION

We propose DeepChunk to leverage deep Q-learning to make chunk-based cache update decisions on the fly in Data Processing Networks, and to jointly optimize both network performance and data processing objectives, including the



(a) Chunk hit ratio (b) Processing stall time (c) Object download time

Fig. 7. Reward breakdown for different policies in Wifi connection. Cache capability is 80 chunks. Weight factors are set to 10 (hit ratio), 1 (stall time), and 0.5 (download time).



(a) Chunk hit ratio (b) Processing stall time (c) Object download time

Fig. 8. Reward breakdown for different policies with 4G connection. Cache capability is 80 chunks. Weight factors are set to 10 (hit ratio), 1 (stall time), and 0.5 (download time).

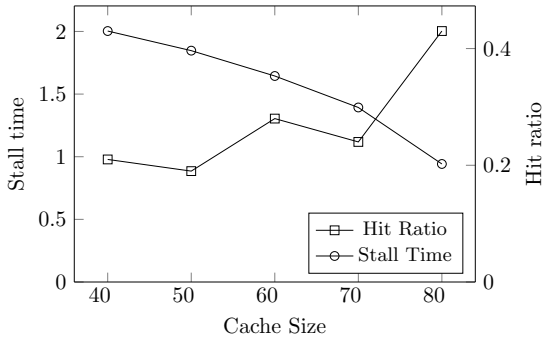


Fig. 9. LRU performance affected by cache size.

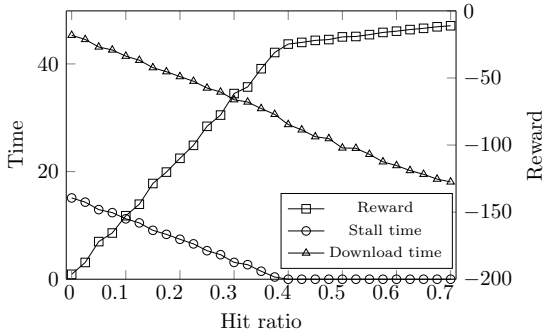
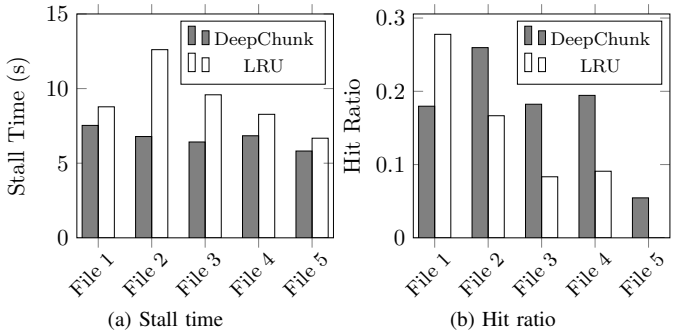
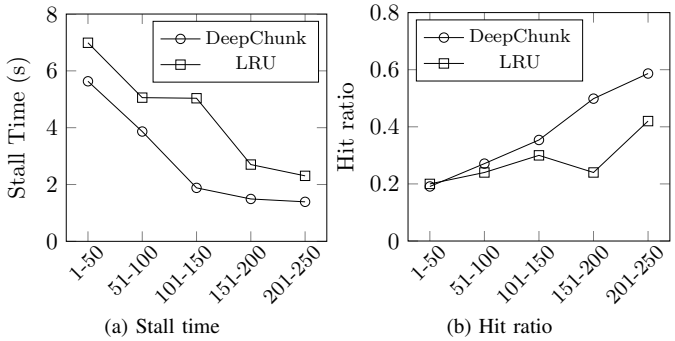


Fig. 10. Reinforcement learning mechanism of DeepChunk decision making.



(a) Stall time (b) Hit ratio

Fig. 11. Reward breakdown for different files. The standard deviations of stall time are 0.628 for DeepChunk and 2.189 for LRU. The standard deviations of chunk hit ratio are 0.074 for DeepChunk and 0.104 for LRU.



(a) Stall time (b) Hit ratio

Fig. 12. Reward breakdown comparison for time bins, when the popularity changes for every 50 requests.

chunk hit ratio, processing stall time, and object download time. Our prototype using Ceph demonstrates significant improvement, i.e., 52.20% in total reward and 68.10% in processing stall time, over a number of baseline caching policies, as well as DeepChunk's ability to adapt to time-varying workload and network conditions. As a future research, we plan to incorporate other metrics such as the Age of Information [44] into DeepChunk and investigate the performance of DeepChunk in a setting with network of caches.

REFERENCES

[1] M. D. de Assunção, A. D. S. Veith, and R. Buyya, "Distributed data stream processing and edge computing: A survey on resource elasticity and future directions," *J. Network and Computer Applications*, vol. 103, pp. 1–17, 2018.

[2] T. Li, Z. Xu, J. Tang, and Y. Wang, "Model-free control for distributed stream data processing using deep reinforcement learning," *Proceedings of the VLDB Endowment*, vol. 11, no. 6, pp. 705–718, 2018.

[3] A. Elgabli, V. Aggarwal, S. Hao, F. Qian, and S. Sen, "Lbp: Robust rate adaptation algorithm for svc video streaming," *IEEE/ACM Transactions on Networking*, vol. 26, no. 4, pp. 1633–1645, 2018.

[4] D. Jiang, B. C. Ooi, L. Shi, and S. Wu, "The performance of mapreduce: An in-depth study," *Proceedings of the VLDB Endowment*, vol. 3.

[5] N. Alliance, "5g white paper," *Next generation mobile networks, white paper*, pp. 1–125, 2015.

[6] Youtube help. https://support.google.com/youtube/answer/1722171.

[7] Y. Kim and I. Yeom, "Performance analysis of in-network caching for content-centric networking," *Comput. Netw.*, vol. 57, no. 13, pp. 2465–2482, Sep. 2013.

[8] D. K. Krishnappa, S. Khemmarat, L. Gao, and M. Zink, "On the feasibility of prefetching and caching for online tv services: a measurement study on hulu," in *International Conference on Passive and Active Network Measurement*. Springer, 2011, pp. 72–80.

[9] D. Shasha and T. Johnson, "2q: A low overhead high performance buffer management replacement algoritm," in *VLDB*, 1994, pp. 439–450.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[11] Ceph. https://ceph.com/.

[12] B. Hubert *et al.*, "Linux advanced routing & traffic control howto," *Netherlabs BV*, vol. 1, 2002.

[13] J. van der Hooft, S. Petrangeli, T. Wauters, R. Huysegems, P. R. Alface, T. Bostoen, and F. De Turck, "HTTP/2-Based Adaptive Streaming of HEVC Video Over 4G/LTE Networks," *IEEE Communications Letters*, vol. 20, no. 11, pp. 2177–2180, 2016.

[14] H. Ahlehagh and S. Dey, "Hierarchical video caching in wireless cloud: Approaches and algorithms," in *ICC 2012*. IEEE, 2012, pp. 7082–7087.

[15] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5g wireless networks," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 82–89, 2014.

[16] S. Sivasubramanian, G. Pierre, M. Van Steen, and G. Alonso, "Analysis of caching and replication strategies for web applications," *IEEE Internet Computing*, vol. 11, no. 1, 2007.

[17] Y. Zhao, J. Wu, and C. Liu, "Dache: A data aware caching for big-data applications using the mapreduce framework," *Tsinghua science and technology*, vol. 19, no. 1, pp. 39–50, 2014.

[18] V. Aggarwal, Y.-F. R. Chen, T. Lan, and Y. Xiang, "Sprout: A functional caching approach to minimize service latency in erasure-coded storage," *IEEE/ACM Transactions on Networking*, vol. 25, no. 6, pp. 3683–3694, 2017.

[19] T. Luo, V. Aggarwal, and B. Peleato, "Coded caching with distributed storage," *IEEE Transactions on Information Theory*, pp. 1–1, 2019.

[20] A. Al-Abbasi, V. Aggarwal, T. Lan, Y. Xiang, M.-R. Ra, and Y.-F. Chen, "Fasttrack: Minimizing stalls for cdn-based over-the-top video streaming systems," *IEEE Transactions on Cloud Computing*, 2019.

[21] A. O. Al-Abbasi, V. Aggarwal, and M.-R. Ra, "Multi-tier caching analysis in cdn-based over-the-top video streaming systems," *IEEE/ACM Transactions on Networking (TON)*, vol. 27, no. 2, pp. 835–847, 2019.

[22] H. Che, Y. Tung, and Z. Wang, "Hierarchical web caching systems: Modeling, design and experimental results," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 7, pp. 1305–1314, 2002.

[23] E. Friedlander and V. Aggarwal, "Generalization of lru cache replacement policy with applications to video streaming," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems (TOMPECS)*, vol. 4, no. 3, p. 18, 2019.

[24] D. S. Berger, R. K. Sitaraman, and M. Harchol-Balter, "Adaptsize: Orchestrating the hot object memory cache in a content delivery network." in *NSDI*, 2017, pp. 483–498.

[25] V. Martina, M. Garetto, and E. Leonardi, "A unified approach to the performance analysis of caching systems," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 2040–2048.

[26] H. Ahlehagh and S. Dey, "Video-aware scheduling and caching in the radio access network," *IEEE/ACM Transactions on Networking (TON)*, vol. 22, no. 5, pp. 1444–1462, 2014.

[27] N. Beckmann, H. Chen, and A. Cidon, "{LHD}: Improving cache hit rate by maximizing hit density," in *15th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 18)*, 2018, pp. 389–403.

[28] Z. Xu, J. Tang, J. Meng, W. Zhang, Y. Wang, C. H. Liu, and D. Yang, "Experience-driven networking: A deep reinforcement learning based approach," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1871–1879.

[29] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *SIGCOMM*. ACM, 2017, pp. 197–210.

[30] G. Alnwaimi, S. Vahid, and K. Moessner, "Dynamic heterogeneous learning games for opportunistic access in lte-based macro/femtocell deployments," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 2294–2308, 2015.

[31] A. O. Al-Abbasi, A. Ghosh, and V. Aggarwal, "Deeppool: Distributed model-free algorithm for ride-sharing using deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2019.

[32] O. Onireti, A. Zoha, J. Moysen, A. Imran, L. Giupponi, M. A. Imran, and A. Abu-Dayya, "A cell outage management framework for dense heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 4, pp. 2097–2113, 2016.

[33] R. Fagin, "Asymptotic miss ratios over independent references," *Journal of Computer and System Sciences*, vol. 14, no. 2, pp. 222–250, 1977.

[34] M. Garetto, E. Leonardi, and V. Martina, "A unified approach to the performance analysis of caching systems," *ACM TOMPECS*, vol. 1, no. 3, p. 12, 2016.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[36] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Eleventh annual conference of the international speech communication association*, 2010.

[37] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *International Conference on Machine Learning*, 2016, pp. 1995–2003.

[38] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Thirtieth AAAI conference on artificial intelligence*, 2016.

[39] M. Hausknecht and P. Stone, "Deep recurrent q-learning for partially observable mdps," in *2015 AAAI Fall Symposium Series*, 2015.

[40] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[41] Tornado web server. http://www.tornadoweb.org/en/stable/.

[42] L. A. Adamic and B. A. Huberman, "Zipf's law and the internet." *Glottometrics*, vol. 3, no. 1, pp. 143–150, 2002.

[43] A. V. Aho, P. J. Denning, and J. D. Ullman, "Principles of optimal page replacement," *JACM*, vol. 18, no. 1, pp. 80–93, 1971.

[44] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *INFOCOM, 2012 Proceedings IEEE*. IEEE, 2012, pp. 2731–2735.