Theme Article: Neural-Enhanced Disaggregated Storage

Efficient Disaggregated Cloud Storage for Cold Videos with Neural Enhancement

Jinyeong Lim, *KAIST, Daejeon, 34141, South Korea*
Juncheol Ye, *KAIST, Daejeon, 34141, South Korea*
Jaehong Kim, *KAIST, Daejeon, 34141, South Korea*
Hwijoon Lim, *KAIST, Daejeon, 34141, South Korea*
Hyunho Yeo, *KAIST, Daejeon, 34141, South Korea*
Junhyeok Jang, *KAIST, Daejeon, 34141, South Korea*
Myoungsoo Jung, *KAIST and Panmnesia, Inc., Daejeon, 34141, South Korea*
Dongsu Han, *KAIST, Daejeon, 34141, South Korea*

Abstract—The rapid growth of video-sharing platforms has driven immense storage demands, with disaggregated cloud storage emerging as a scalable and reliable solution. However, the proportional cost of cloud storage relative to capacity and duration limits the cost-efficiency for managing large-scale video data. This is particularly critical for cold videos, which constitute the majority of video data but are accessed infrequently. To address this challenge, this paper proposes Neural Cloud Storage (NCS), leveraging content-aware super-resolution (SR) powered by deep neural networks. By reducing the resolution of cold videos, NCS decreases file sizes while preserving perceptual quality. optimizing the cost trade-offs in multi-tiered disaggregated storage. This approach extends the cost-efficiency benefits to a greater range of cold videos and achieves up to a 21.2% reduction in total cost of ownership (TCO), providing a scalable, cost-effective solution for video storage.

Video sharing services are playing a pivotal role in today's entertainment, social networks, and business. While videos contain diverse information such as visual and audial contents, it require significantly large storage capacity to accommodate such information. For instance, YouTube needs 2.64 TB of additional storage capacity every minute, handling 500 hours of video uploads.[1] Similarly, Netflix manages 2 Petabytes of data for its video service.[2]

In response to such increasing storage capacity demand, cloud storage has emerged as a scalable and reliable solution. Based on disaggregated storage system,[3] it offers storage capacity as much as users need by connecting multiple storage resources through high-bandwidth networks such as RoCE or Infiniband.

Digital Object Identifier 10.1109/XXX.0000.0000000

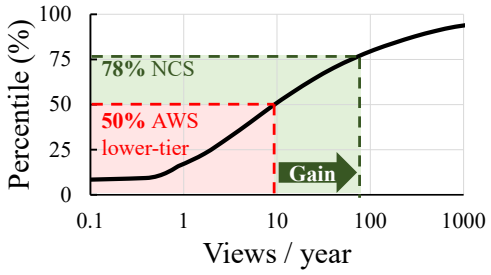


FIGURE 1. NCS expands the cold video coverage that can benefit from multi-tiered service provided by AWS.

In addition, it offers reliable services by redundantly storing the videos across geographically distributed locations.

Even though the disaggregated cloud storage can scale to accommodate petabytes of videos, cost-efficient management of such data poses another chal-

This article has been accepted for publication in IEEE Micro. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MM.2025.3562625

lenge to video sharing services. Since cloud service providers often charge per-GB fees for user's data, the cost of storing the petabyte-scale videos is significant and continues to grow as new videos are uploaded. To reduce the storage cost, one might think of leveraging a highly skewed access frequency (popularity) of videos to establish a multi-tiered storage system. Specifically, a small percentage of videos account for the vast majority of views, while cold videos with fewer views occupy most of the storage capacity. For example, the most popular 3.67% of videos on YouTube account for 93.61% of total views while 65.44% of videos are considered cold, having 100 or fewer views.[4] By managing such cold videos in cheaper storage tier, we can cut off the storage cost without significantly harming the overall performance.

However, existing multi-tiered cloud storage services can reduce storage costs only for a limited number of low-access videos. For example, AWS S3's "Infrequent" tier cuts storage costs by nearly half ($0.021 → $0.0125/GB*month) compared to the "Frequent" tier but imposes a significant retrieval cost of $0.01/GB per access. As a result, only videos with an average of 0.85 views or fewer views per month can benefit from such services. Based on our analysis of historical YouTube view data, as shown in Figure 1, only half of all videos could achieve storage cost savings.

To address this issue, this paper proposes a solution by leveraging neural enhancement, specifically content-aware super-resolution (SR), to improve the cost efficiency of frozen video cloud storage without compromising video quality. Although additional computational cost is incurred during inference, this approach is suitable for frozen videos with low access frequency, typically accessed once per month. Based on this, we propose a prototype system called Neural Cloud Storage (NCS), offering the following benefits: First, it reduces storage costs by lowering the resolution of frozen videos and optimizes the trade-off between storage and retrieval costs, thereby reducing the total cost of ownership (TCO). Second, by reducing TCO, it extends the range of frozen videos benefiting from multi-tier services from 50% to 78%, supporting videos accessed up to 91 times annually, as illustrated in Figure 1. Finally, NCS has the potential to reduce TCO by 21.2% compared to storing cold videos on the cheapest AWS storage.

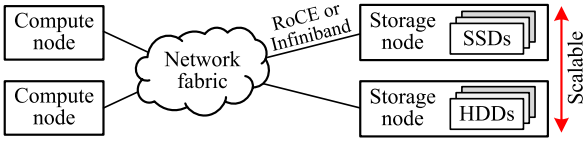


FIGURE 2. Disaggregated storage architecture in modern cloud.

Background

Disaggregated cloud storage

Figure 2 shows the typical architecture of disaggregated storage in modern cloud environments. At a high level, disaggregated storage comprises multiple storage nodes connected to compute nodes via a high-bandwidth network fabric. The storage node employs an array of storage devices (e.g., SSDs, hard disks) which can be accessed through either a storage protocol such as NVMe-over-Fabric or a distributed file system such as Lustre. Compute node can secure the storage capacity as much as users need (e.g., for handling video uploads) by connecting the storage node as needed.

In such a system, there are two types of costs that contribute to the total cost of ownership (TCO) for storage: (1) storage cost and (2) retrieval cost. Storage cost refers to the cost of storing video data, while retrieval cost refers to the cost of retrieving the video. The costs are calculated by considering that storing and retrieving consume power. In addition, it also considers that storing and retrieving consume the lifetime of the storage device. For example, in an SSD, storing data directly consumes the device's lifetime because the underlying media have a limited number of write/erase cycles they can endure. It also requires internal tasks to maintain the data without bit errors, consuming both the lifetime and power.

To this end, the TCO of the cloud video storage system can be formulated as follows:

$$\begin{aligned}\text{TCO}(V_i) = [&\text{Storage cost}(V_i) \times \mathit{time}_i + \\ &\text{Retrieval cost}(V_i) \times \mathit{access}_i]\end{aligned}$$

$$\text{TCO of video cloud storage} = \sum_{i=1}^{N} \text{TCO}(V_i)$$

where the unit of Storage cost$(\cdot)$ is $\frac{\$}{\text{time}}$; while the unit of Retrieval cost$(\cdot)$ is $\frac{\$}{\#\text{access}}$;

V_i is the i-th video stored in the cloud; time_i is the stored time of V_i and access_i is the number of access during time_i, respectively.

According to the formula, the TCO per video can be minimized by reducing storage or retrieval costs for cold videos, leading to a lower overall TCO for video cloud storage. Multi-tier storage in the cloud offers trade-offs between storage and retrieval costs based on access frequency. By selecting an appropriate storage tier based on video access frequency, TCO can be minimized. For instance, calculations indicate that if a video is accessed fewer than 11 times per year, AWS's Infrequent tier is more cost-effective than other storage options.

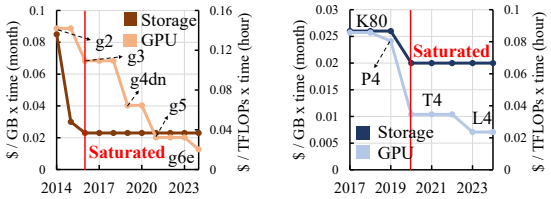
Content-aware super-resolution

Super-resolution (SR) is a technique that transforms low-resolution (LR) images into high-resolution (HR) counterparts. With the advent of deep neural networks (DNNs), SR has achieved impressive results in reconstructing LR images. Recent research has highlighted that SR, when applied with a content-aware approach, can substantially reduce Internet bandwidth consumption in video streaming.[5] This method involves customizing SR DNN models for individual video content—referred to as content-aware SR, which offers superior enhancement compared to DNNs trained on generic datasets. In this study, we explore the application of content-aware SR in cloud storage systems, demonstrating its potential to reduce storage costs, especially for cold video content.

Motivation

Why is neural cloud storage promising?

Neural-enhancement, particularly content-aware SR, shows great potential for cold video cloud storage for three main reasons. First, content-aware SR reduces storage costs while preserving video quality. For example, downscaling a 20-minute 4K video to FHD decreases file size from 1.17GB to 0.77GB, cutting storage costs by 34%. While this process incurs inference costs, they are minimal for long-term stored videos like cold videos. Second, cloud storage prices have stabilized, while cloud computing costs continue to decline. As shown in Figure 3(a), storage prices remain steady, whereas computing instance costs are dropping. This trend makes neural cloud storage an increasingly cost-effective solution. Lastly, content-aware SR typically requires training for each video, leading to high computational costs. However, clustering similar videos and training a single DNN per group efficiently distributes these costs. Most videos can be processed using one DNN trained on a representative video, reducing training costs as the cluster grows. This approach significantly lowers expenses for large-scale video-sharing platforms.



(a) Amazon Web Services (b) Google Cloud Platform

FIGURE 3. The price trend of storage and GPU instance of cloud providers

More flexible storage options for cold videos

Efficient storage management is essential for both cost reduction and improving user experience. Beyond cost, one of the key distinctions between storage tiers is their availability. Storage tiers with lower availability often result in long tail latency and reduced reliability, leading to degraded service quality and higher data recovery costs in certain scenarios. To address this, NCS provides storage tier upgrades to improve availability without additional cost. This ensures better accessibility and user experience, even for data such as surveillance footage that may occasionally require rapid access.

How to maximize the benefits of neural cloud storage?

To maximize the advantages of neural cloud storage, there are opportunities for further optimization, particularly in two key areas. First, while content-aware SR methods significantly reduce storage costs by storing downscaled video content, further improvements are possible. Videos uploaded to the cloud are typically stored in a compressed format using conventional codecs that are not tailored for SR applications. Developing an SR-optimized encoding approach for these codecs could further enhance storage efficiency. Second, while infrequent access to cold videos reduces retrieval frequency, the computational cost of SR inference remains significant. By leveraging the distinct properties of video data, it is feasible to reduce overall inference costs without compromising video quality. This study explores these two areas of improvement and their practical impact on reducing TCO.

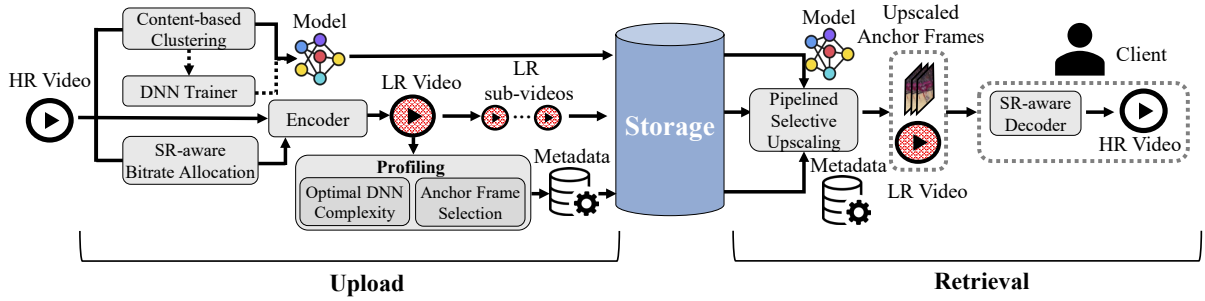


FIGURE 4. High-level overview of Neural Cloud Storage

Design of NCS

Overview

Figure 4 presents an overview of NCS architecture. NCS minimizes storage costs by dynamically adjusting video resolutions and storage tiers based on popularity. Cold videos—those with an annual average view count of 91 or fewer—are stored at lower resolutions in the cheapest storage tier. When requested, NCS restores these videos to their original resolution using content-aware SR DNNs and metadata. To further reduce TCO, NCS optimizes storage by profiling videos and embedding metadata alongside compressed files.

Producing Content-Aware SR DNN

Training a content-aware SR DNN for each video is computationally expensive, making large-scale deployment impractical. To mitigate this, NCS clusters videos based on content similarity, allowing SR DNNs to be shared within groups. This reduces training costs while preserving video-specific enhancements. Unlike methods that rely solely on metadata classification or generic SR models, NCS employs a two-stage clustering process. It first classifies videos using metadata from uploaders or platforms. Since metadata alone lacks precision, NCS refines clustering with a vision encoder[6] that extracts feature vectors from key frames, ensuring more accurate grouping. NCS also dynamically adjusts clusters. If a video does not fit existing clusters, a new one is created with an initial SR DNN. As more videos join, the model undergoes iterative fine-tuning, continuously improving restoration quality. This adaptive approach optimizes computational efficiency while maintaining high-quality results, outperforming static clustering and generic SR models.

Profiling Video for TCO Optimization

NCS reduces storage costs through low-resolution compression and further lowers TCO with additional optimizations. These include refining video encoding and minimizing SR DNN computing costs, enabled by profiling processes performed prior to storage.

SR-aware bitrate allocation. Conventional video encoders prioritize visual quality but frequently neglect SR quality. NCS's analysis reveals that the relationship between allocated bitrate and SR quality varies across frames based on unique content characteristics. Building on this, NCS proposes a novel SR-aware video encoding method that dynamically allocates bitrate based on each frame's SR quality potential. To achieve this, the video is first encoded at multiple target bitrates, and SR quality is evaluated for key frames within each GOP (Group of Pictures) instead of all frames to reduce computational complexity. This data is used to determine the optimal bitrate allocation, minimizing overall video size while preserving target SR quality. A greedy algorithm is utilized to efficiently profile SR quality. Initially, each GOP is assigned the minimum bitrate. The bitrate is iteratively increased for the GOP offering the highest SR quality gain per unit of additional bitrate until the desired SR quality is achieved. This approach strikes an effective balance between computational efficiency and quality optimization.

Optimizing DNN complexity. Similar to the relationship between bitrate and SR quality, the sensitivity of SR quality to DNN complexity also varies across frames. Utilizing this, NCS dynamically adjusts DNN complexity for each frame to minimize SR overhead while maintaining overall SR quality. Similar to SR-aware bitrate allocation, NCS profiles SR quality across various DNN configurations. NCS assesses the SR quality of key frames within GOPs and adjusts DNN complexity based on FLOPS. Subsequently, a greedy algorithm is used to determine the optimal DNN complexity for each frame that satisfies the target quality. This adaptive strategy ensures efficient utilization of computational resources by focusing on frames where

increased complexity results in the most significant quality improvement.

Anchor frame selection. To reduce SR computing costs, NCS adopts an anchor frame selection strategy.[7] This approach uses specific anchor frames to effectively support the upscaling of other frames. Information from these anchor frames, upscaled using DNNs, is leveraged to upscale the remaining frames through an SR-integrated decoder.[7] This method drastically reduces the frames requiring direct DNN inference, cutting computational costs while preserving acceptable quality.

Once the appropriate SR DNN is selected by clustering, the video is encoded with an optimized bitrate allocation. The encoded video is then analyzed to determine the optimal DNN complexity, which is used to assign a DNN to each chunk and select the corresponding anchor frames. Due to the dependencies between each optimization step, the process follows a fixed order. This information is stored as metadata with the video. The metadata, being minimal (<100KB), has a negligible impact on storage requirements.

Video Restoration

NCS encounters two challenges in SR-based video restoration: delays during the SR process, which degrade user experience, and re-encoding of restored frames, which introduces significant computational overhead and increases processing time. To address these challenges, NCS leverages pipelining and hybrid video encoding.

Pipelining with sub-videos. NCS employs a pipelining mechanism that strategically overlaps SR processing with video restoration to minimize additional delays. By segmenting videos into sub-videos along the temporal axis, NCS enables concurrent data retrieval and SR restoration. To maximize pipelining efficiency, NCS follows two key principles when dividing videos. First, to eliminate re-encoding and preserve video quality during merging, segmentation is performed strictly at the GOP level. Second, while shorter sub-videos enhance pipeline throughput, excessively small segments introduce inefficiencies due to the inherent latency threshold in storage retrieval. Balancing these factors, NCS optimally determines sub-video sizes to ensure efficient storage and processing. During restoration, sub-videos are processed in parallel through the pipeline and seamlessly reassembled into a single video before transmission. This design not only minimizes restoration latency but also maintains high video quality, achieving an optimal trade-off between speed and fidelity.

Hybrid video encoding. To reduce re-encoding overhead, NCS adopts a hybrid video encoding approach.[7] This method efficiently minimizes computational overhead by restoring only anchor frames to their original resolution using SR DNN. The restored anchor frames are losslessly compressed with an image codec and packaged with the stored LR video into a single file using a hybrid encoder, then transmitted to the client. During playback, the anchor frames and SR-aware integrated decoder restore and decode the remaining non-anchor frames to their original resolution. This approach is hundreds of times more efficient than re-encoding the entire video with conventional encoders.

Evaluation

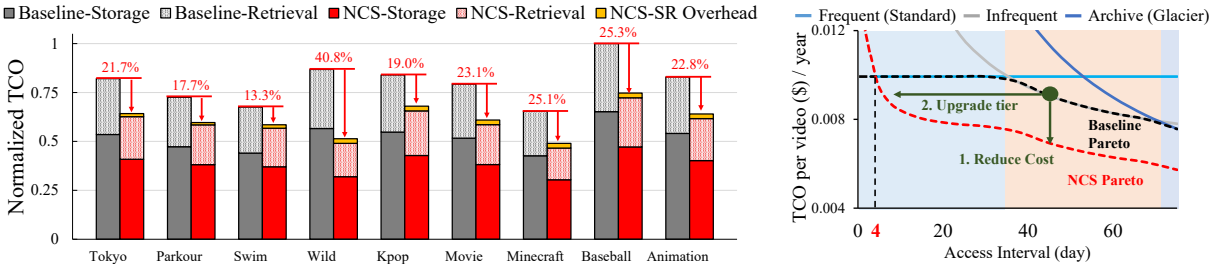
Experimental Setup

Setup. We use a modified version of EDSR[8] model for content-aware SR. To reduce costs, we employ a lightweight (~100KB) DNN, pre-trained on general image datasets and fine-tuned per video. TensorRT enables fast inference. Videos are encoded with VP9 and FFMPEG, using YouTube's encoding settings except for the target bitrate. HR and LR videos share the same encoding configurations, except for resolution. The dataset consists of various 2160p YouTube videos, each one minute long and centered on specific content themes. For SR-aware bitrate allocation and optimal DNN complexity, a greedy search algorithm considers 10 candidate bitrates and DNN complexities. About 10–15% of the frames are selected as anchor frames. Clustering and profiling are performed once, incurring negligible costs over long-term storage and are excluded from TCO calculations.

Baseline. We use AWS S3 multi-tier storage as storage nodes and AWS EC2 g6e.xlarge spot instances (0.185$/hour) as compute nodes. The baseline stores the HR-encoded video in the most cost-effective storage option based on access patterns, referred to as "Baseline Pareto".

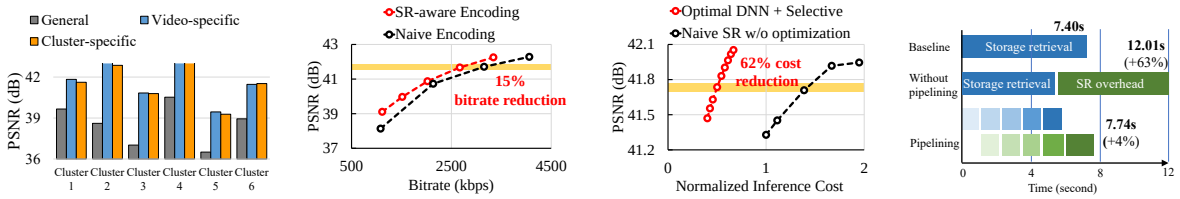
Cost-benefit Analysis

In this section, we perform a cost-benefit analysis of integrating our optimization techniques into the existing cloud storage system. Given practical limitations, we did not implement the full system end-to-end to directly measure costs. Instead, we estimated and compared



(a) Cost comparison with different contents (Access interval = 45 days) (b) NCS pareto vs. Baseline pareto

FIGURE 5. NCS provides better TCO for cold video (access interval > 4 days) compared to Baseline.



(a) (b) (c) (d)

FIGURE 6. Component analysis (a) Clustering (b) SR-aware encoding (c) SR-aware decoding (d) Clustering

the TCO by aggregating the costs of individual components.

Figure 5(a) shows the normalized TCO comparison for storing various video content between the Baseline and NCS over a 45-day access interval. Unlike the Baseline, NCS incurs SR overhead costs; however, these costs become negligible when the access interval is short. The evaluation shows that NCS reduces TCO by up to 40.8% compared to the baseline, depending on the video content. This indicates that while SR performance varies with the content and dynamic characteristics of videos, NCS consistently delivers high efficiency across various scenarios.

Figure 5(b) illustrates the Pareto optimality between NCS and the Baseline, demonstrating that the selected storage tier varies based on the access interval. NCS achieves a lower TCO than the Baseline for all cold videos with an access interval of 4 days or more. Additionally, as the popularity of the video decreases, the TCO savings ratio increases, highlighting that NCS is particularly optimized for storing cold data. Based on this calculation, Figure 1 shows that NCS can reduce the TCO for 78% of all videos in YouTube and achieve a 21.1% reduction in TCO compared to the Baseline.

Optimizing Storage Costs and Tier Allocation

As shown in Figure 5(b), NCS allows a lower TCO than the baseline or access to a higher storage tier at the same TCO for video storage. Based on this calculation, NCS offers the opportunity to upgrade 16% of YouTube videos to an upper storage tier. This demonstrates that NCS not only reduces TCO but also enhances user experience through upgraded storage tiers.

Component Analysis

Clustering. To validate the proposed clustering method, vision encoder was used to extract feature vectors from 40 sports-themed videos, resulting in six clusters with 4–8 videos each. Figure 6(a) compares SR quality using cluster-specific DNNs, a general DNN, and video-specific DNNs. Cluster-specific DNNs outperformed the general model and achieved similar quality to video-specific DNNs. This demonstrates that content-based clustering reduces the number of DNNs needed while maintaining high performance and computational efficiency.

SR-aware encoding. Figure 6(b) compares the rate-distortion performance between naive encoding and SR-aware encoding with SR-aware bitrate allocation. The quality of SR is measured using PSNR, while the size of video is represented by the bitrate. The curve clearly highlights the advantages of SR-aware encoding, showing that it can achieve up to a 15% reduction in bitrate while maintaining the same SR quality.

SR-aware decoding. Figure 6(c) compares the SR-aware decoding approach, which leverages optimized

DNN complexity and anchor-frame selection, with naive SR methods in terms of quality and computational efficiency. As shown in the figure, SR-aware decoding can achieve the same SR quality while reducing inference costs by an average of 62%.

Restoration. Figure 6(d) illustrates the latency comparison of pipelining in NCS. The experiment utilized a 10-minute 4K video file with a size of 703 MB. NCS benefits from reduced storage retrieval latency due to smaller file sizes, but without pipelining, the overall retrieval time increases by 63% compared to the baseline due to the time required for SR processing. By employing pipelining with a minimum sub-video size of 100 MB, the overall latency is reduced by up to 36%, achieving a latency level similar to the baseline.

Future work & Discussion

In this section, we explore the possibilities for further expanding and refining our work.

Temporal popularity-aware storage strategy

NCS optimizes video storage costs by analyzing the annual average access interval of videos. However, this approach assumes that video popularity remains constant over time, which is not realistic. Most videos receive the majority of their views within days of upload, while only a small fraction maintain long-term viewership.[9] A more effective strategy is to store videos in their original resolution upon upload and transition them to cold storage once their view count drops below a threshold, reducing costs dynamically. Integrating a temporal popularity prediction model with NCS would enable proactive storage management, further optimizing cost savings across a larger proportion of videos.

Adaptive popularity-based storage strategy

Most videos uploaded to video platforms exhibit a logarithmic viewing pattern over time.[9] However, some videos may unexpectedly surge in views long after upload, leading to frequent lower-tier storage access and raise TCO. To address this, these videos can be moved to upper-tier storage for LR-to-HR conversion, though this may degrade perceptual quality and affect the original viewing experience. A potential solution to this problem is to implement exception handling for videos with highly volatile viewing patterns. Specifically, the HR versions of these videos can be backed up in archive storage (e.g., AWS S3 Deep Archive) and moved to upper-tier storage if the video experiences a sudden surge in views. his approach minimizes TCO spikes for videos with volatile viewing patterns while maintaining high-quality playback and preserving the user experience.

Conclusion

We propose Neural Cloud Storage (NCS) as a cost-effective solution for storing cold videos in disaggregated cloud storage systems. By leveraging neural enhancement, particularly content-aware super-resolution, NCS demonstrates potential benefits such as cost reduction, storage tier upgrades, and expanded applicability to cold videos. With the continued advancement of deep learning and computational resources, we believe neural enhancement will play a pivotal role in revolutionizing disaggregated cloud storage systems, making them more efficient and cost-effective.

Acknowledgements

This article is a full-length version of an earlier work.[10] We appreciate anonymous reviewers. This work was supported by Samsung Electronics Co., Ltd [IO221107-03428-01] , the National Research Foundation of Korea (NRF) [RS-2024-00340099], and the Institute for Information and Communications Technology Promotion (IITP) [RS-2023-00221040].

REFERENCES

1. "Google Transparency Report, [Online]." Available: https://transparencyreport.google.com/?hl=en.
2. J. Summers, T. Brecht, D. Eager, and A. Gutarin, "Characterizing the workload of a netflix streaming video server," in 2016 IEEE International Symposium on Workload Characterization, pp. 1–12, IEEE, 2016.
3. W. Bai, S. S. Abdeen, A. Agrawal, K. K. Attre, P. Bahl, A. Bhagat, G. Bhaskara, T. Brokhman, L. Cao, A. Cheema, et al., "Empowering azure storage with RDMA," in 20th USENIX Symposium on Networked Systems Design and Implementation, pp. 49–67, 2023.
4. R. McGrady, K. Zheng, R. Curran, J. Baumgartner, and E. Zuckerman, "Dialing for videos: A random sample of youtube," Journal of Quantitative Description: Digital Media, vol. 3, 2023.
5. H. Yeo, Y. Jung, J. Kim, J. Shin, and D. Han, "Neural adaptive content-aware internet video delivery," in 13th USENIX Symposium on Operating Systems Design and Implementation, pp. 645–661, 2018.

6. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning, pp. 8748–8763, PMLR, 2021.
7. H. Yeo, H. Lim, J. Kim, Y. Jung, J. Ye, and D. Han, "Neuroscaler: Neural video enhancement at scale," in Proceedings of the ACM SIGCOMM 2022 Conference, pp. 795–811, 2022.
8. B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp. 136–144, 2017.
9. L. C. Miranda, R. L. Santos, and A. H. Laender, "Characterizing video access patterns in mainstream media portals," in Proceedings of the 22nd international conference on world wide web, pp. 1085–1092, 2013.
10. J. Lim, J. Ye, J. Kim, H. Lim, H. Yeo, and D. Han, "Neural cloud storage: Innovative cloud storage solution for cold video," in Proceedings of the 15th ACM Workshop on Hot Topics in Storage and File Systems, pp. 1–7, 2023.

JINYEONG LIM is currently working toward a Ph.D. degree at KAIST, Daejeon, 34141, South Korea.. His research interests include computer systems, video processing, and machine learning optimization. Contact him at jylim9999@kaist.ac.kr

JUNCHEOL YE is a currently working toward a Ph.D. degree at KAIST, Daejeon, 34141, South Korea. His research interests include machine learning system, video system, and network system. Ye received his M.S. degree from KAIST. Contact him at juncheol@kaist.ac.kr.

JAEHONG KIM is a researcher affiliated with KAIST, Daejeon, 34141, South Korea.. His research focuses on immersive multimedia systems, networked computer systems, and AI-driven video streaming. Kim received his Ph.D. degree in Electrical Engineering from KAIST. Contact him at jaehong950305@gmail.com.

HWIJOON LIM is a researcher affiliated with KAIST, Daejeon, 34141, South Korea. His research focuses on improving machine learning and networking systems by designing efficient algorithms and implementing system-level optimizations in real-world systems. Lim received his Ph.D. degree in Electrical Engineering from KAIST. Contact him at hwijoon.lim@gmail.com.

HYUNHO YEO is a researcher affiliated with KAIST, Daejeon, 34141, South Korea. His research focuses on optimizing video streaming and networking systems through deep learning. Yeo received his Ph.D. degree in Electrical Engineering from KAIST, Daejeon, South Korea. Contact him at chaos5958@gmail.com.

JUNHYEOK JANG is currently working toward a Ph.D. degree at KAIST. His research interests include CXL, disaggregated systems, and machine learning acceleration. Contact him at jhjang@camelab.org

MYOUNGSOO JUNG is a full professor in the School of Electrical Engineering, KAIST. He is also the CEO of Panmnesia, inc. His research interests include computer architecture, operating system, storage systems, non-volatile memory, parallel processing, heterogeneous computing, and CXL. Contact him at m.jung@kaist.ac.kr

DONGSU HAN is a full professor in the School of Electrical Engineering and Graduate School of Artificial Intelligence, KAIST. His research interest include networked/cloud systems design, AI for systems, and systems for AI. He is a corresponding author of this article. Contact him at dhan.ee@kaist.ac.kr