

Errata for “The Impact of Highly Interactive Workloads on Video-on-Demand Systems”

Andrew Brampton B.Sc (Hons)

1 General Changes

The Background work considered should be more critically discussed. In particular with respect to instantaneous popularity of all content items as an issue separate from popularity development over time.	
After	<p>In Section 2.2.1.1 Popularity, it was made clear that Zipf can not be used over time, by adding the following paragraph into the middle of the section: “The fact that Zipf did not fit well over a year time scale is often overlooked and the reasons for this are typically misunderstood. Zipfian models are useful for static distribution, those which do not have a temporal component. For example, Zipf was first developed by George Kingsley Zipf when studying the frequency of words appearing in a corpus of natural language. The corpus would not change over time, <i>i.e.</i> words would not be added or removed. This differs from popularity of videos over a timescale as it is common for new videos to be added and removed over time, and for the popularity of the videos to change over time. Instead the Zipf models should be used to model the popularity over a shorter periods, such as daily or weekly, or used to blindly model the rank of objects on daily bases. For example, objects on a daily bases may follow a Zipf distribution, but instead of noting the popularity of each object, instead note the popularity of each rank. This will then more accurate model the expected popularity on any given day, and is more useful for the design of caching systems.”</p>

<p>Model fitting: the different models should be discussed with respect to their use case (i.e. consider carefully what they have been intended for). They should then be applied considering their suitability to express a specific case investigated within the thesis.</p>	
Note	<p>Firstly I've created two pages of new content, 4.1 Probability Distribution, which outlines the main models and what they are useful for. I have also gone through the evaluation and added some extra text where needed to help explain how the model is relevant.</p>
Before	<p>In the last paragraph of Section 4.3: "Additionally, the skewed nature of this distribution is most likely because it is impossible to have a negative seek value."</p>
After	<p>"Regarding long-ranged seeks, the log-normally distributed models imply that some very large distance seeks do occur, but the majority of seeks are shorter. Additionally, the skewed nature of this distribution is most likely because it is impossible to have a negative seek value."</p>
Before	<p>Second/Third paragraph of Section 4.4: "Our analysis reveals that object popularity does not follow the typical power-law distribution observed within CDNs [CWVL01, AKEV01, YZZZ06], but instead is a normal distribution with parameters $\mu = 60$ and $\sigma = 32$. This can be attributed to the nature of our videos and the relatively few new objects each day."</p>
After	<p>"Typically object popularity with CDNs and VoD systems follows a power-law distribution [CWVL01, AKEV01, YZZZ06], however, our analysis reveals otherwise. Instead the ranking of objects best fitted a normal distribution with parameters $\mu = 60$ and $\sigma = 32$. There are two reasons that power-law was not the best fit. Firstly, the catalogue of 88 videos was not very large, and secondly, power-law distributions do not fit well if the objects are constantly changing. Instead power-law fits better if a snapshot of rank <i>vs.</i> popularity is taken each day and aggregated."</p>
Before	<p>Third/Fourth paragraph of Section 4.4 "The popularity of one-second segments for all the videos exhibit a Weibull..."</p>
After	<p>"Again, the popularity of one-second segments might be best suited by a power-law, however Zipf and Pareto did not fit well. Instead, the popularity of one-second segments for all the videos exhibit a Weibull..."</p>
Note	<p>It was commented by Griwodz that perhaps I remove Table 4.3 (the table that lists how well different models fitted each metric). If you feel my use of models is better without the table then I will remove it, otherwise I'd like to keep it.</p>

<p>E.g. Zipf expresses popularity only for periods where popularity does not significantly change, whereas this is the case in the presented work.</p>	
Note	<p>Hopefully the clarification of Zipf in section Section 2.2.1.1 should suffice. Also, in my evaluation Zipf was not discussed as a good fit for any distribution, including popularity. Instead popularity of objects was normally distributed, and popularity of segments was log-normal (however, the popularity of segments did not change over time, therefore it shouldn't be impacted).</p>

Explain fittings and elaborate what you try to achieve in this thesis (Zipf, R-Square, Kolmogorov-Smirnov text, etc.)	
Note	The newly added Section 4.1, Probability Distribution, and more specifically Section 4.1.1 helps explain why fitting was used, and what I tried to achieve.

Elaborate more on interactivity and user issues such as:	
Impact of assumed trick-modes, (e.g. FFW, Rewind, time-code controls, etc.)	
Note	Section 4.2 outlines how often each trick-mode was used, along with how this influenced the users. Perhaps more is needed? One change I have made is to actually use the term “trick-mode”.
After	This was added in Section 3.1 paragraph 5. “We also wanted the interface to offer modern interactive controls (sometimes called <i>trick-modes</i>), such as the ability to seek forward and backwards, as well as pausing and resuming.”
User behaviour and the influence of bookmarks.	
Note	The influence that the bookmarks had on the users was already summarised in paragraph 1 and 4 of Section 4.11. Also the fact that users seeked from bookmark to bookmark (as explained in Section 4.8) also highlights this fact. However, if you feel I should explicitly write more about this then I’ll add it to Section 4.11.
Content semantics and their impact, E.g. Sports vs. music videos	
Note	I have added a couple of paragraphs into Section 4.11 (Summary).
After	<p>“Though out the experiment different genres of videos were available to the users, namely sporting and musical videos. Only the analysis of the “Argentina vs. Serbia and Montenegro” football match and the “Eurovision song contest” were shown in this chapter, however other sporting events were available on the site such as Formula 1 racing, International Cricket, and other miscellaneous recordings of music channels. Similar patterns were observed for each video, however, semantics of the content did have some impact of how the users consumed the data.</p> <p>All videos exhibited similar patterns, for example, popularity was generally centered around the bookmarked segments, and that the viewing duration was far shorter than the full length of the video. However, minor differences were found, for example, the music channels had greater variance in the popularity of each bookmark (which were placed at the beginning of individual music videos). This can easily be attributed to users only being interested in particular artists or videos, whereas viewers of sporting events would be interested in every highlight (and therefore every bookmark). Similar differences were found in the inter-seek times, session times, and hotspot lengths, as the semantics of the content would determine how long particular hotspots were. However, metrics such as the number of interactions, or bookmark longevity stayed the same, as these did not appear to be directly impacted by the content.”</p>

2 Checklist Changes

Page 2: too general remark on multicast, please revise	
Before	“For example, network or application-level multicast is no longer suitable under heavy loads of interactivity”
After	“For example, conventional network and application-level multicast is not suitable for providing interactivity.”

Page 6, §2.1 - causality sentence, popularity and causality, please revise	
Before	“This has been driven by incredible demand, causing these video-on-demand (VoD) applications becoming increasingly popular.”
After	“This has been driven by incredible demand, causing new video-on-demand (VoD) systems to appear, almost daily, to serve different niches.”

Page 17, push deadline discussion to be clarified	
Before	“With tree based distribution, it is typical for push distribution to be used. As each peer will require the segments at a similar deadline...”
After	“Push distribution is typically used for live streaming in combination with tree based distribution. In live streaming each peer will require the segments by a similar deadline...”

Page 29: deepen discussion on popularity change over time. This was observed but no appropriate model. Discuss influence on your own model	
After	<p>Add at the end of the section 2.2.1.3 “Although popularity of objects change over time, it appears these changes are very specific to the viewing population and genres of the objects. In the current research no single model has been found which accurately explain the observed behaviours, however, it is clear that this is an important metric for cache design.”</p> <p>Also section 4.5, ‘Longevity’, has been adapted to discuss how change in popularity effects my own models.</p>

Page 31, §2.2.1.4 2nd paragraph: explain environment and significance	
Before	“For example, Chesire <i>et al.</i> showed that 85% of all sessions lasted less than 5 minutes with a median session duration of 2.2 minutes”
After	<p>“For example, in 2001, whilst studying streaming traffic on a large university campus, Chesire <i>et al.</i> showed that 85% of all sessions lasted less than 5 minutes with a median session duration of 2.2 minutes”.</p> <p>The significance of these findings is explained in the 5th paragraph of this section.</p>

Page 35, §2.2.2.2: Clarify distribution of session duration	
Note	The shape of the distribution is discussed, however my notes ask about the “top 3%” comment
Before	“These results were similar to Chesire <i>et al.</i> who found a similar long-tailed distribution with their top 3% of sessions being more than a hour in length.”
After	“These results were similar to Chesire <i>et al.</i> who found a similar long-tailed distribution with their top 3% of the population being more than a hour in length.”

Page 37: Clarify sentence starting with “However”, second paragraph with respect to the long-tail content	
Before	Two sentences have been added onto the end of the 2nd paragraph to help clarify.
After	“For example, one caching policy can be used for most popular content, whilst another can be used for the more niche content. As such, the niche caching policy may only store the content in local caches for a short, whereas the very popular content is kept available for a longer period of time, on a more global scale.”

Page 38: Pre-fix caching, reference appropriate pre-fix caching paper (one of the later papers that e.g. are introducing delayed harmonic broadcasting by Paris).	
Before	“Techniques such as prefix caching can also aid in deciding which segments are the most useful to cache and replicate.”
After	“Techniques such as prefix caching[SRT99, HNG+99] can also aid in deciding which segments are the most useful to cache and replicate.”

Page 39: “perhaps encourage”, please rephrase	
Before	“This system was designed to provide controls which allowed and perhaps encouraged the use of interactivity.”
After	“This system was designed to provide powerful, yet simple interactivity controls, which would hopefully encourage more interaction between the users and the system.”

Page 42: state the rules of describing metadata and also web 2.0 about tagging.	
Note	The rules for the metadata was content specific and is described later in section 3.3, so therefore I added a forward reference.
Before	“This metadata included the title and description of the video as well as marking the location of key events within the videos which would become <i>bookmarks</i> .”
After	“This metadata included the title and description of the video as well as marking the location of key events within the videos which would become <i>bookmarks</i> (more details on what was bookmarked is listed in Section 3.3).”

Page 47: standard deviation of 30, please remove from text	
Before	“In total there were 88 videos, each video on average was 2.5 hours in length with a standard deviation of 30 minutes.”
After	“In total there were 88 videos, with an average length of 2.5 hours”

Page 49: explain the 383 effect of being the most clicked on (without being really viewed)	
Note	A extra paragraph has been added to explain this
After	“An observant reader will note that the most popular video had 383 unique users, yet the analysis is concentrated on <i>arg-scg</i> and <i>eurovision</i> with only 123-131 unique users. The reason for this is that the most popular content, a collection of ‘cheesy’ music videos, had a very short average session duration. The video was the newest content on the site for many weeks, as such was at the top of the list of videos. We speculate that newcomers to the site would click on this video to understand what the site had to offer, but quickly stop. Shortly afterwards they would continue to explore the other videos on the site, which were perhaps better to their liking. These shorts views were therefore not representative of a typical viewing session and thus the analysis does not concentrate on them.”

Page 50, paragraph 2: the comment on short session duration needs to be verified;	
Note	I double checked the literature, and the facts I stated were correct

Page 50, paragraph 1: explain how small and long seeks are treated (are small seeks joined?)	
Note	I changed the wording to make it clear these are individual seeks, when using the seek buttons.
Before	“Small forward seeks were used a combined 24.9% of the time, whereas backward seeking was only used 7.67%. These actions only accounted for the relatively small seeks (10, 30, and 60 seconds)”
After	“Small individual forward seeks were used a combined 24.9% of the time, whereas individual backward seeking was only used 7.67%. These actions only accounted for the short-seeks buttons (10, 30, and 60 seconds)”

Page 52: Clarify figure 4.3 a and b’s relationship (probably in figure caption)	
Before	“Figure 4.2: CDF of seek distance: (a) Small scale, (b) Large scale”
After	“Figure 4.2: CDF of backward and forward seek distances: (a) All seek (b) All seeks cropped at 200 seconds”

Page 54: clarify Eurovision bookmark placement	
Before	“The vertical lines signify the position of the bookmarks; note for the <i>eurovision</i> video there were no bookmarks after 6000 seconds since the performances bookmarked were only in the first half.”
After	“The vertical lines signify the position of the bookmarks; note for the <i>eurovision</i> video there were no bookmarks after 6000 seconds as only the performances were bookmarked and they all appeared in the first half of the video.”

Page 54: second but last and last sentence on this page, please clarify	
Before	“We have seen that bookmarked videos provide a content format with specific segments of interest (goals, for example).”
After	“We have seen that bookmarks within videos cause segments of high interest and popularity, for example, goals within a sporting event.”

Page 55: clarify figure 4.6 description (resp. caption)	
Before	“Bookmark utilisation over time, following initial usage”
After	“Bookmark utilisation within all videos over time, following initial usage”

Page 55: clarify the popularity statement (“popularity within a video”)	
Before	“The segments which were popular when the video was first published were still popular within the video weeks later, long after the video had lost popularity.”
After	“For example, the segments which were popular within that video when it was first published were still popular within the video weeks later, long after the video had lost it overall popularity.”

Page 57: Clarify Poisson and Pareto statement	
Note	I noticed I made a mistake here, the two distribution were Weibull and Pareto, not Poisson.
Before	“For educational content, inter-see times have also been shown to be <i>Poisson</i> or <i>Pareto</i> distributed [AKEV01]. We however found only two thirds of our videos had inter-see times that could be suitably modelled by a Pareto distribution, and none that could be modelled well with a Poisson distribution.”
After	“For long educational content, inter-see times have also been shown to be Weibull distributed or a combination of Weibull for the body and Pareto for the tail [AKEV01]. We found that most of our videos had inter-see times that could be suitably modelled by a Weibull distribution, and two thirds which could be modelled with Pareto alone.”

Page 77/78: mention ageing alongside flushing	
Note	Added new paragraph after the flushing discussion.
After	“A final suggested technique would use the concept of aging, whereby, entries in the table who have not be observed recently are removed from the table. This technique may provide the most relevant entries at the cost of additional overhead for each entry.”

Page 78, 5.3: correct that some pull based systems have structure, e.g. PROMISE, DAGSTREAM	
Note	Changed the sentence, and references to these systems are provided in a earlier chapter
Before	“The pull based system does not try to configure the peers into any structured way, instead creating a mesh of connections between the peers.”
After	“The pull based system do not typically try to configure the peers into any structured way, instead creating random meshes of connections between the peers.”

Page 80: include alternatives to greedy discussion, e.g. controlled multicast (Gao & Towsley , ICMCS 1999; Eager, Vernon, Zahorjan, MIS 1999, Griwodz, Liepert, Zink, Steinmetz, PER 2000)	
Note	Added a new paragraph in the middle of the patching discussion. I have only referenced “Gao & Towsley , ICMCS 1999” and “Griwodz <i>et al.</i> PER 2000” here. The reference “Eager, Vernon, Zahorjan, MIS 1999” was added to the concluding paragraph.
After	“A middle ground is controlled multicast, which adds access controls limiting how many patching channels can be created. For example, the controlled <i>CIWP</i> algorithm [GT99] uses a mathematically optimal scheduling algorithm that limits the rate at which patching channels are created. Another technique, Lambda Patching [GLZS00], allows the server to decide on the patching window sizes, based on currently observed popularity and interarrival times.”

Page 82: reconsider last sentence of 5.3.2.1 with respect to bandwidth in periodic broadcast	
Note	I’m unsure of the issue here, and my notes don’t help.

Page 86, figure 5.7: mention uplink capacity in the caption	
Before	Instead of changing the caption I have expanded the paragraph which talks about the experiments. “Content node(s) (those sourcing the media) are attached to the aforementioned transit node(s), whereas clients are attached to randomly selected members of the stub domain(s).”
After	“Content node(s) (those sourcing the media) are attached to the aforementioned transit node(s), whereas clients are attached to randomly selected members of the stub domain(s). The content nodes are connected via highly provisioned links, whereas, the clients are limited to a typical asymmetric link (1Mbit down, 256kb up).”

3 Spelling/Grammar Changes

§2.1.5 Para. 6, Sentence 9 - Added word	
Before	“One problem for all trees, is that they may very deep, causing a high latency (or lag) for the peers near the bottom.”
After	“One problem for all trees, is that they may become very deep, causing a high latency (or lag) for the peers near the bottom.”

§2.2.1.1 Para. 2, Sentence 1 - Duplicate reference	
Before	“Dan et al. in 1994 [DSS94a]”
After	“Dan et al. in 1994 [DSS94]”

§5.1 Para. 11, Sentence 2 - Added a comma	
Before	“we have potentially stopped video being transferred which might have normally been skipped over”
After	“we have potentially stopped video being transferred, which might have normally been skipped over”

§5.2.2 Para. 1, Sentence 1 - Fixed spelling of sparsely	
Before	“The workloads observed in this thesis exhibited sparely distributed”
After	“The workloads observed in this thesis exhibited sparsely distributed”

§5.2.3 Para. 3, Sentence 4 - Fixed grammar	
Before	“it has been purposely been designed so”
After	“it has purposely been designed so”

§5.3.1 Para. 4, Sentence 4 - Added missing word	
Before	“ After 10 minutes they would have caught up with the channel and thus can disconnect from their behind channel.”
After	“ After 10 minutes they would have caught up with the ahead channel and thus can disconnect from their behind channel.”

4 Other Changes

Title Page - Changed title	
Before	“The Impact of Highly Interactive Workloads on Multimedia Systems”
After	“The Impact of Highly Interactive Workloads on Video-on-Demand Systems”

Page 2: too general remark on multicast, please revise	
Before	“For example, network or application-level multicast is no longer suitable under heavy loads of interactivity”
After	“For example, conventional network and application-level multicast is not suitable for providing interactivity.”

§3.1 Para. 2 - Expanded on what metadata	
Before	“Administrators would then manually add metadata to the system describing the files as well as marking the location of key events within the videos which would become <i>bookmarks</i> .”
After	“Administrators would then manually add metadata to the system describing the files. This metadata included the title and description of the video as well as marking the location of key events within the videos which would become <i>bookmarks</i> .”

Missing reference to Figure 2.1	
Before	Added to section 2.1.1 “The HTTP servers used for streaming are typically hosted by a content distribution network (CDN).”
After	“The HTTP servers used for streaming are typically hosted by a content distribution network (CDN) as depicted in Figure 2.1”

Missing reference to Figure 2.2	
Before	Added to section 2.1.3 “The Sky+ PVR, for example, records broadcast TV received via a satellite dish.”
After	“The Sky+ PVR, for example, records broadcast TV received via a satellite dish as seen in Figure 2.2”

Missing reference to Figure 2.4	
Note	I removed this figure, there was no way to easily fit it into the text. I did however really like this image :(