Summary for Ofcom

Technologies for distributing linear content over IP

5 May 2023

797048496-185b

Multicast is the most technically efficient technology for IP-delivery of linear content, but the market may evolve to rely on 'best effort' unicast delivery from CDNs

- The delivery of 'linear' content, including live TV and other scheduled content, will be increasingly via IP[1] networks. This move is being driven by a need for greater quality, functionality and interactivity, as the market responds to competition from hyperscale streamers
- We define linear content as: *Audiovisual content, including TV content and content from online providers, which is shown at a scheduled time. This can include 'live' events, such as sports, comedy, entertainment and other public interest events, as well as the scheduled showings of pre-recorded content. The nature of the content is such that many viewers will watch the same content at the same time*
- There are three basic techniques for delivering linear content over IP: unicast, multicast and 'deep CDN'-assisted[2] unicast. If a peak linear viewing event of ~30 million viewers (~15 million households) was delivered entirely over IP, we estimate that network traffic would more than double on backhaul links for 'deep CDN'-assisted delivery, though multicast brings minimal impact
- While multicast appears to be an efficient technical solution in principle, various factors are very likely to constrain its impact, including the need to 'opportunistically' switch between multicast and unicast delivery, the need for end-to-end control within an ISP's[3] network, and the future prevalence of linear vs. non-linear viewing
- We can see at least two possible future scenarios:
  - One scenario where multicast (from selected ISPs) sits alongside CDN-assisted unicast. In this case, it may be possible for ISPs to compete for a share of PSBs' [4] spend on content delivery with their multicast solution
  - An alternative scenario where the market continues to rely on 'best effort' deep CDN-assisted unicast, because most internet content is provided from hyperscalers, which are showing little interest in non-unicast solutions such as multicast
- Several accompanying issues are also relevant to Ofcom's thinking in the potential move to IP delivery of linear content:
  - the energy consumption of the equipment needed to serve the content is small compared to energy consumption of general internet connectivity and in-home devices, and there are various techniques available to minimise the energy impact of content servers
  - some important aspects of the end-user experience would benefit from optimisation, including latency (delay)
  - finally, some policy issues would need to be tackled to support a transition, including coverage, reliability, take-up, along with clarifications of net neutrality constraints and potentially new obligations on ISPs

[1] IP = internet protocol; [2] CDN = content delivery network; [3] ISP = internet service provider; [4] PSBs = public service broadcasters

This report presents the findings of our research to investigate techniques for delivering linear video content over IP networks

- Ofcom issued a requirement for a report which investigates:
  - the distribution of linear online video content, and
  - how the networks and value chain might evolve to support increasing levels of linear viewing over IP
- A wide range of content types is considered (e.g. sport, news, drama) in a variety of formats (e.g. SD[1]/HD[2]/4K) and functionalities (e.g. targeted advertising, object-based media)
- The scope of work is focused specifically on 'linear' viewing (see definition below); on-demand and user-specific content is considered in the study as part of the wider context of video delivery

Our definition of *linear content*

Audiovisual content, including TV content and content from online providers, which is shown at a scheduled time. This can include 'live' events, such as sports, comedy, entertainment and other public interest events, as well as the scheduled showings of pre-recorded content. The nature of the content is such that many viewers will watch the same content at the same time

- The project assumes a range of market trends will continue:

Start-stop viewing will continue. Users will come to expect to be able to pause, resume and catch up with the programme schedule

On-demand viewing will remain an important (and at times more prevalent) mode of consuming TV and video content

Take-up of higher definitions and quality of content will continue, driven by the natural upgrade cycle of TV devices and competition from a wide range of streaming services

There will be more interactivity and personalisation. This will include interactive viewing experiences, personalised advertising and interactive use of object-based media

The majority of functionality will be performed in the network, and (for example) start-stop viewing will not require a PVR[3]-type device[5]

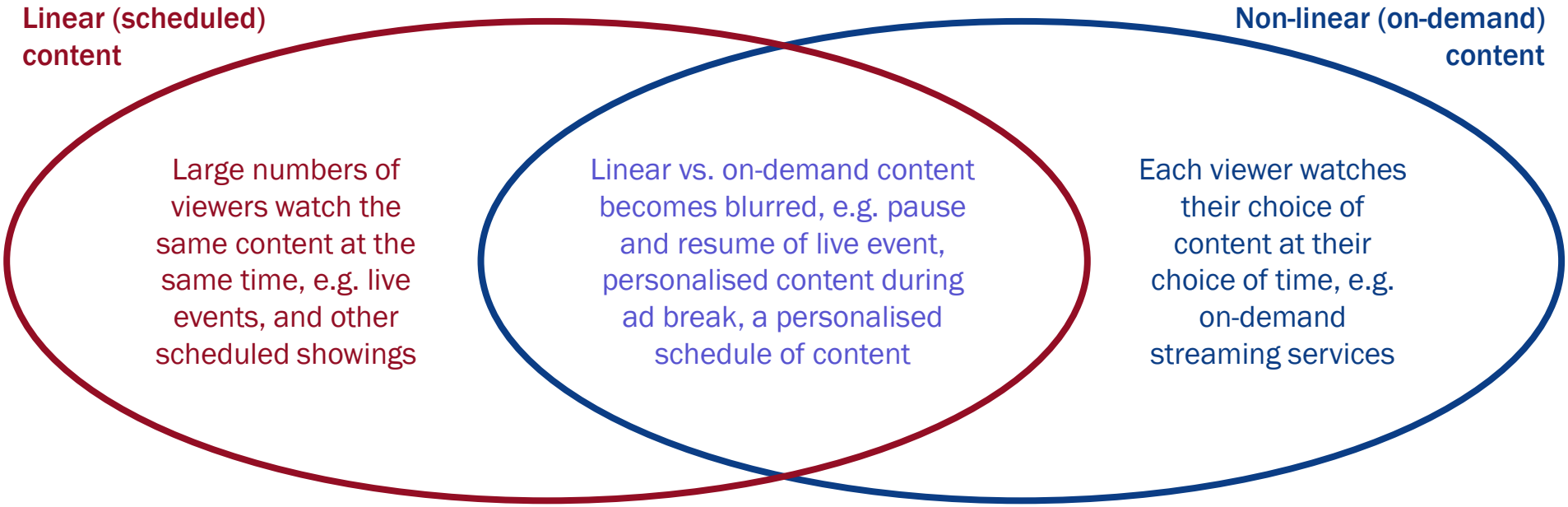
Users should be able to consume their linear content via any ISP[4] (e.g. the ISP they happen to be currently using)

[1] SD = Standard Definition; [2] HD = High Definition; [3] PVR = Personal Video Recorder; [4] ISP = Internet Service Provider

[5] It is possible to pre-load pre-recorded content to in-home storage in advance of a scheduled viewing time, and thus reduce any associated network traffic peaks. However, as this method is not suitable for live events, we do not consider it further

The line between linear and on-demand is becoming more blurred, and while most data may be created by on-demand, peak traffic may come from linear consumption

Overview of the relationship between linear (scheduled) and non-linear (on-demand) content

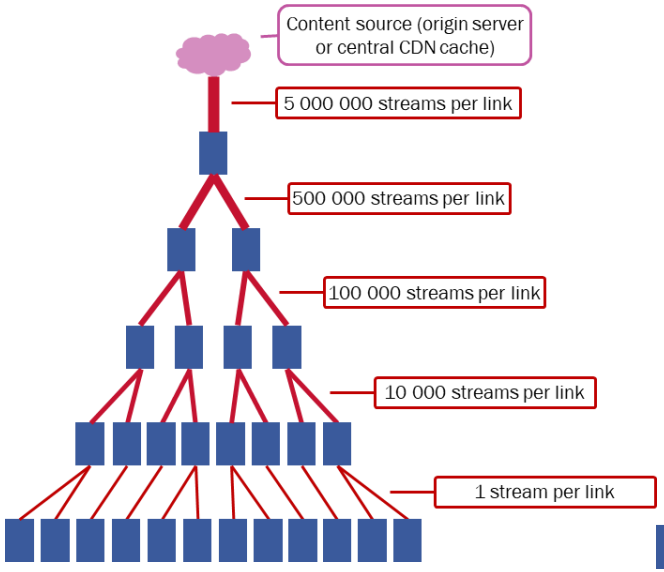


Future proportion of viewing time is uncertain, but likely to create peak traffic demand (in bit/s)

Likely to account for majority of viewing time, and therefore majority of data consumption (in bytes)

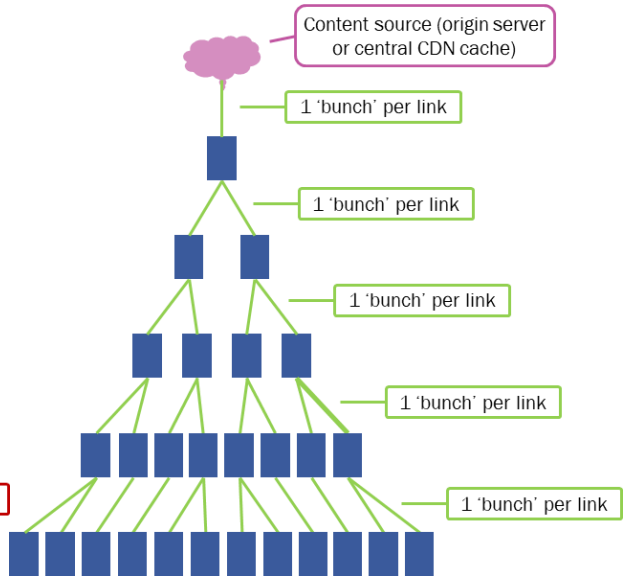
There are three main technical techniques for delivering linear content on IP networks: unicast, multicast and deep CDN-assisted unicast

Unicast content delivery



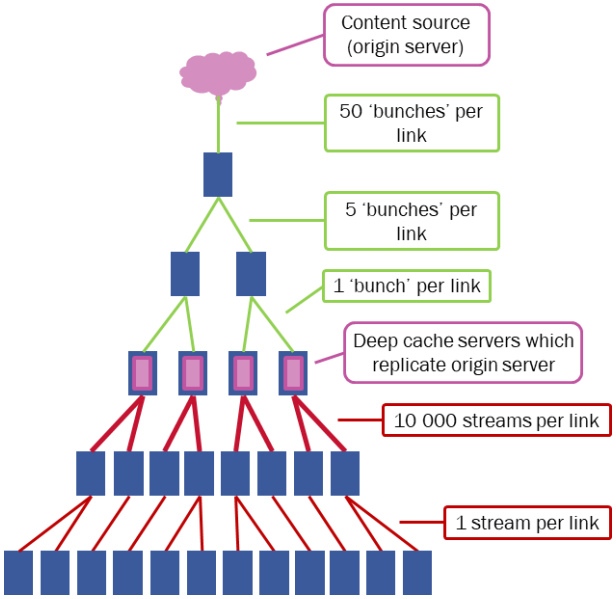
- In unicast video delivery, each user has their own dedicated stream all the way through the network
- This model creates large traffic demands on the higher levels of the network
- The model is technically and commercially impractical for linear viewing at scale: included for reference only

Multicast content delivery



- Where many viewers are watching the same thing at the same time, multicast can be used
- This model transmits one 'bunch' of streams[1] of the content on each link, with each node replicating the stream for the nodes below
- The model is very bandwidth efficient for linear content

Deep CDN-assisted unicast



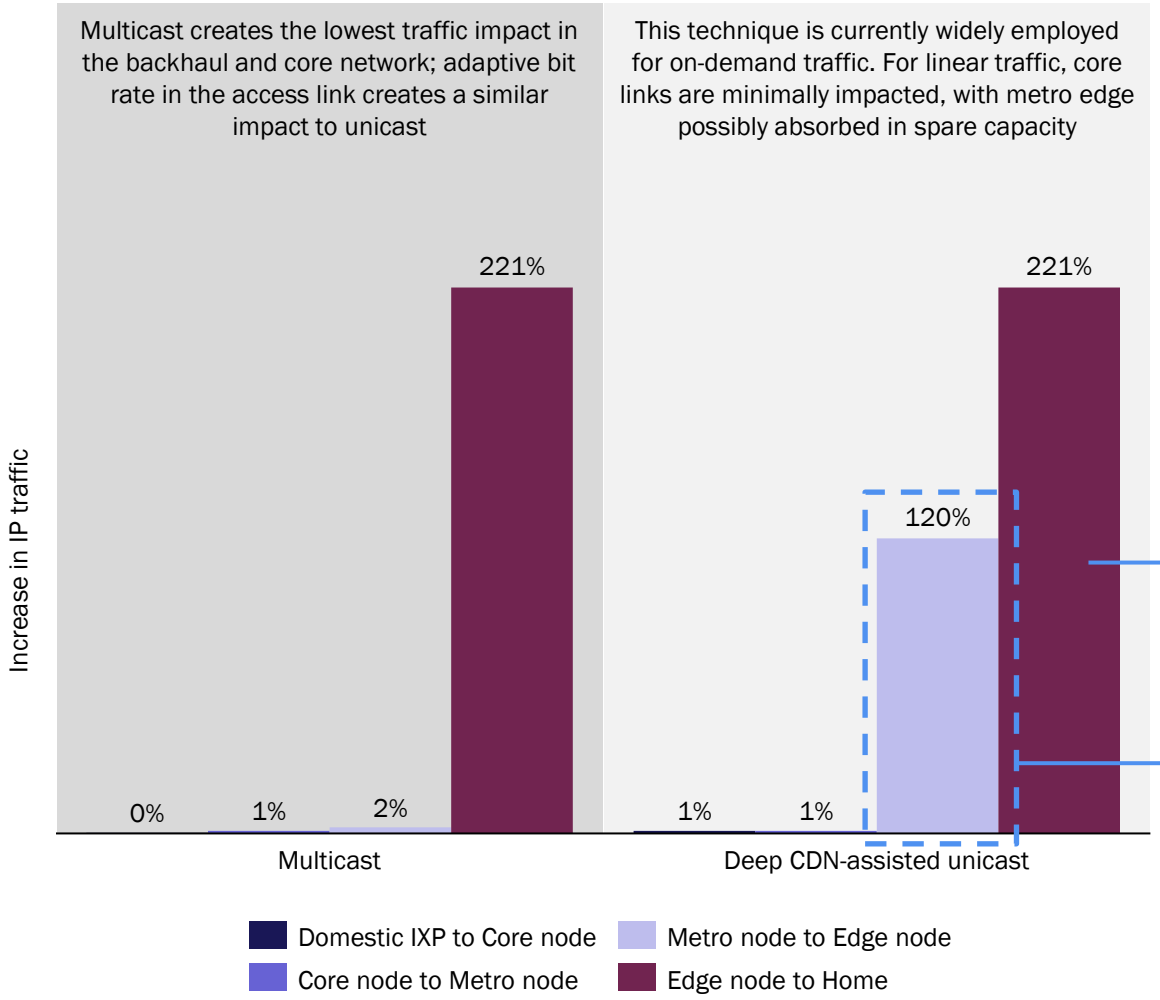
- Most on-demand video content today is delivered via a deep CDN-assisted unicast model
- Cache servers replicate the content close to end users, reducing load on upper links
- The technique can be used for linear content, with significantly lower traffic demands on higher links than unicast

[1] A 'bunch' of streams refers to the small number of parallel versions of a piece of content, which are provided at different levels of quality (e.g. definition, refresh rate, dynamic range) to suit different end-user player capabilities | CDN = Content Delivery Network

analysys mason

A peak IP linear viewing event of 15 million households would more than double the traffic on backhaul links for CDN-based delivery, though multicast brings minimal impact

Estimated increase in IP traffic per link due to additional linear traffic at each network layer, 2030



Key assumptions:

- Network is serving 4.5 million households @ 30% market share (i.e. 15 million households in total are viewing the linear event);
- Multicast streams made available for 70 channels; three video definitions (SD, HD, 4K)
- Results show increase over base traffic, which is an extrapolation of today's total data consumption: predominantly from on-demand viewing, but with a very small amount of linear viewing
- Linear video bandwidth includes a +50% uplift over on-demand content as live content cannot be compressed as effectively as on-demand

Increase in traffic on the access link is assumed to be absorbed in spare capacity of fibre-enabled access networks (a blended average of the assumed take-up of different video definitions shown here)

Traffic on the metro-to-edge (backhaul) link is affected by the number of viewers in the CDN-assisted case

Source: Analysys Mason | [1] IXP = Internet Exchange Point

Although multicast is an elegant solution in theory, in practice various factors may limit its implementation or impact

Limitations on the implementation and impact of multicast

Streaming will switch between unicast and multicast on an *opportunistic* basis

- Linear streams delivered over IP may start in unicast before subsequently switching to multicast:
  - multicast may only be made available for more popular content, on a dynamic basis
  - a unicast stream will allow some players to start more quickly than if relying solely on multicast
- Targeted advertising will be delivered via unicast[1], requiring a switch away from multicast
- Stop-start viewing behaviour will put users onto unicast, until they 'catch back up to schedule'

Switching to unicast, for any of these dynamics, will create an increase in traffic demand, compared to where the delivery had stayed on multicast. Further explanation of the mechanics of opportunistic multicast is given later in the report

ISPs probably require end-to-end control

- Most internet traffic is delivered on an 'open' or 'over-the-top' basis, across multiple network operators. This type of open delivery is unlikely to be possible for multicast because:
  - the large number of standards, and lack of agreed use of an addressing space, means that multicast does not work well across different networks (including non-ISP in-home routers)
  - a degree of traffic prioritisation is likely to be helpful to make multicast work technically and commercially
  - content providers may not want to provide and receive both unicast and multicast streams (i.e. they may prefer to deal only in unicast in the playout of the video stream from their servers, and the receipt of the video stream into their apps/clients)

This dynamic means that multicast may only be implemented end to end by certain large ISPs

The future prevalence of linear viewing is uncertain

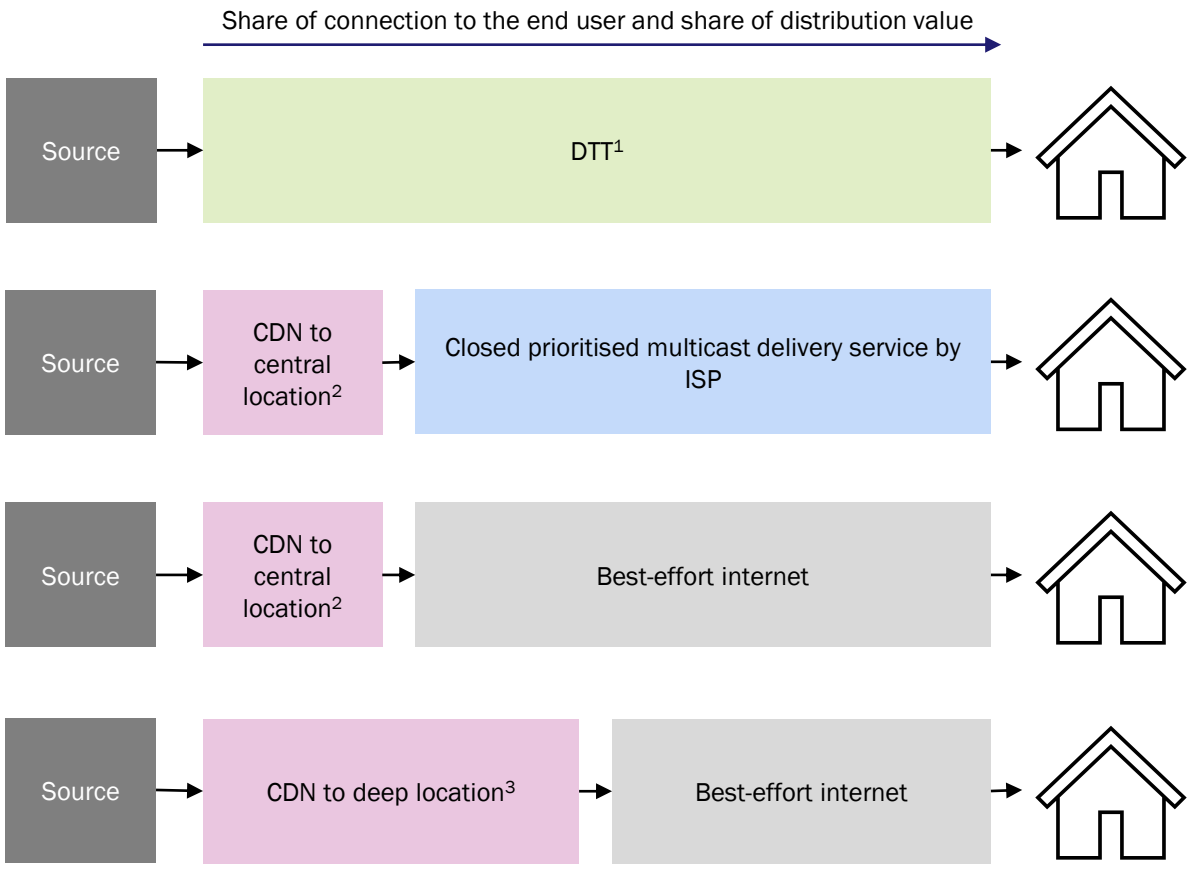
- The preference for linear vs. non-linear (on-demand) viewing is changing, with a shift towards non linear
- While high-profile peaks in network traffic load are currently caused by linear events, non-linear events may also create peaks, e.g. a 'rainy boxing day' with large amounts of on-demand viewing

These dynamics mean that the network may be dimensioned such that 'business as usual' linear peaks can be accommodated from unicast infrastructure. However, major linear events (e.g. royal occasions, sports finals) still have the potential to create the 'peak of peaks', by increasing the number of concurrent viewers

[1] We note that there is a 'continuum' of targeting, ranging from regional down to individual households. Pre-loading of adverts to local storage would also alleviate associated traffic peaks, where storage is available

One possible outcome for the market is that multicast (from selected ISPs) sits alongside CDN-assisted unicast in a mixed delivery scenario

Alternative commercial models for future delivery of linear content



PSBs are currently paying for the service of delivering their content to the end user, via the DTT network,

Some ISPs are considering a competing content delivery service, which would use a multicast solution to deliver the content to the end user. The service would likely interconnect with CDNs at central locations to avoid complex interconnection arrangement with PSBs[4]

The new service would run in parallel to, and share the same infrastructure as, conventional best-effort internet services. The new service would need to be classified as a 'specialised service' under net neutrality rules to enable the ISPs to charge PSBs for the service

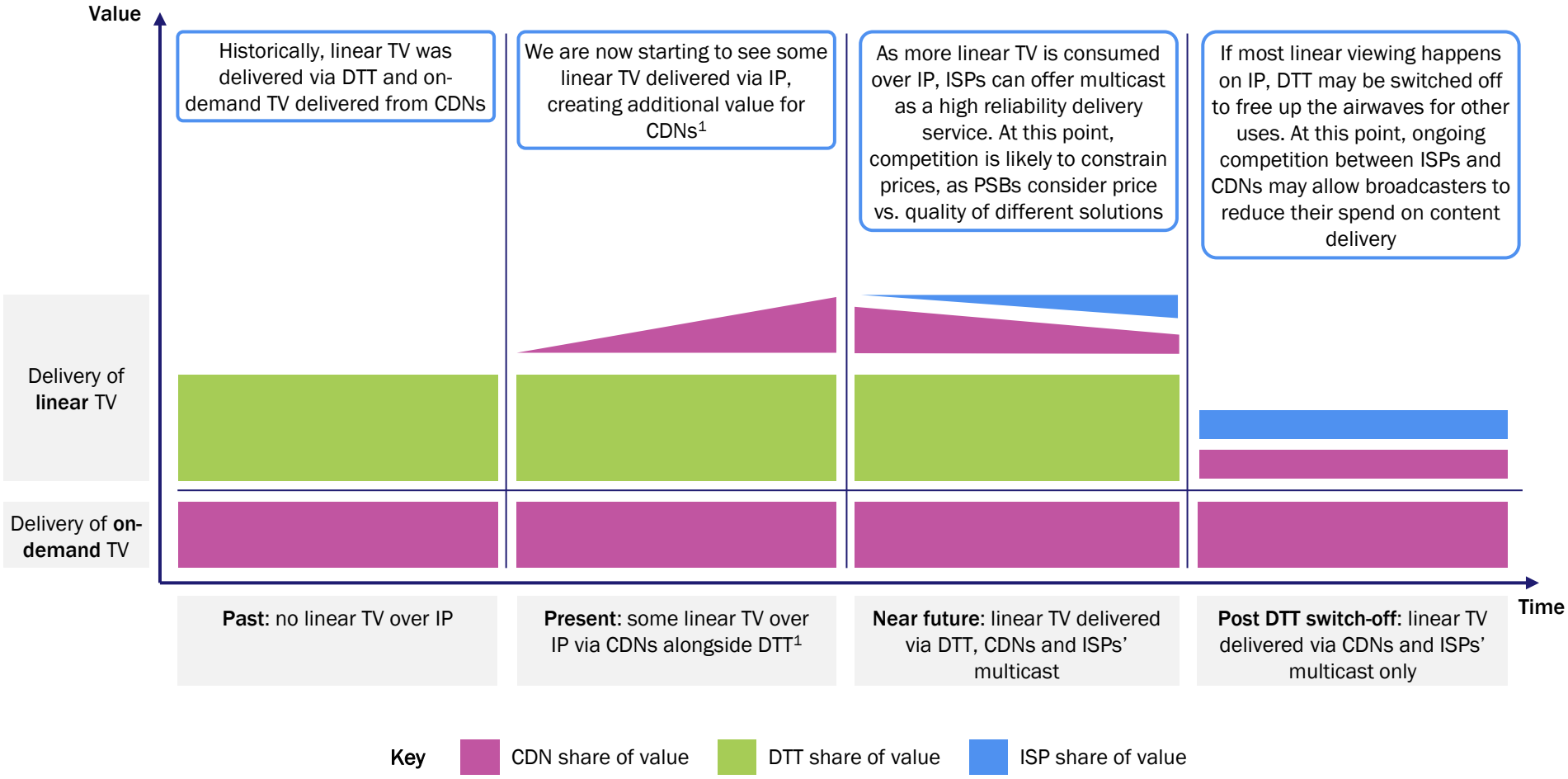
However, an incumbent competitor to the new service would be a best-effort service provided from CDN servers at a deep location. In order to compete, the multicast service must be materially more reliable and/or less expensive than a deep CDN service operating over the least congested part of the network.

[1] CDNs are also used in the DTT model, but not shown here for the purposes of clarity; [2] Central locations could include London, Slough and possibly Manchester; [3] Deep locations include up to around 100 locations in large towns and cities (e.g. 'metro' nodes); [4] PSB = Public Service Broadcaster

In the scenario where ISPs offer multicast, it may be possible for them to compete for a share of PSBs' spend on services for delivering linear content to end users

Illustration of the evolution of share of value in delivery of TV content

Directionally illustrative | Not to scale



[1] The small amount of TV content currently delivered via closed ISP multicast is not shown for clarity

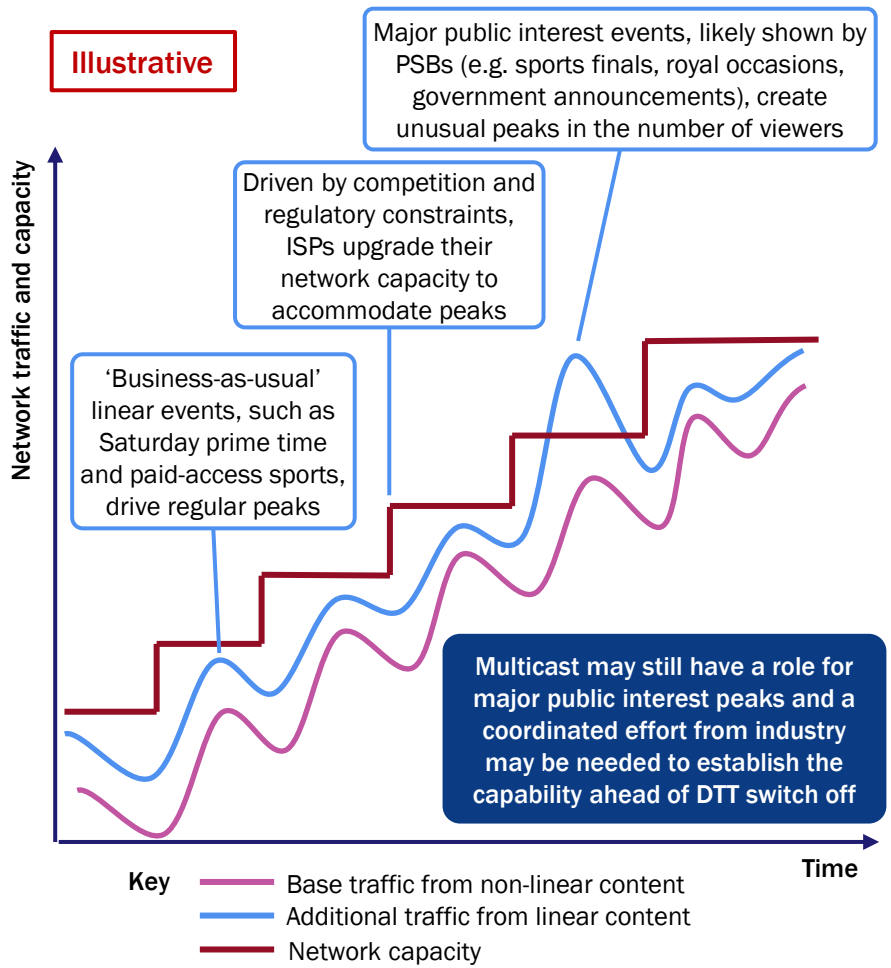
[2] Due to their economies of scope and scale, hyperscalers can potentially offer CDN services at close to marginal cost

analysys mason

Another scenario is that the market continues to rely on 'best-effort' unicast delivery from CDNs, but multicast may still have a role for major public interest peaks

- Most internet traffic (mainly video traffic) is sourced from a small number of 'hyperscale' content and application providers (CAPs), including Google, Amazon, Netflix, and Disney
- To date, hyperscalers have shown limited interest in multicast:
  - linear content may represent a small part of total content
  - targeted advertising may be key to their business case for linear services, which requires unicast anyway
  - CAPs may be waiting for demand to materialise before supporting multicast (a 'chicken and egg' problem)
- This creates a disincentive for smaller providers to also pursue multicast, because:
  - ISPs will continue to upgrade the capacity of their networks, to support 'good enough' delivery of mainly unicast content, because they face competition from other ISPs, and net neutrality rules prevent them from blocking, throttling or charging for the traffic sourced from specific CAPs
  - content providers can make a compelling proposition for deeply locating cache servers for their content in the ISPs' network,[1] because it is likely to save the ISP cost compared to the content being delivered from a more central interconnect or via IP transit
- These dynamics are illustrated in the picture on the right, whereby the network dimensioned for hyperscaler unicast traffic is sufficient to also meet most linear demands

Potential evolution of network capacity and traffic peaks



[1] There are complex strategies around using multiple CDN providers, including sharing peak loads across different CDNs at different times

analysys mason

Various technical and commercial techniques can manage the energy impact of linear IP delivery, though streaming represents a small proportion of the total

- There are various techniques which mean that streaming servers can meet peak traffic requirements, without creating unduly high energy consumption during off-peak times:

Modern server equipment has various techniques for saving power during off-peak times, including deactivating CPU cores and memory modules, and going into sleep mode

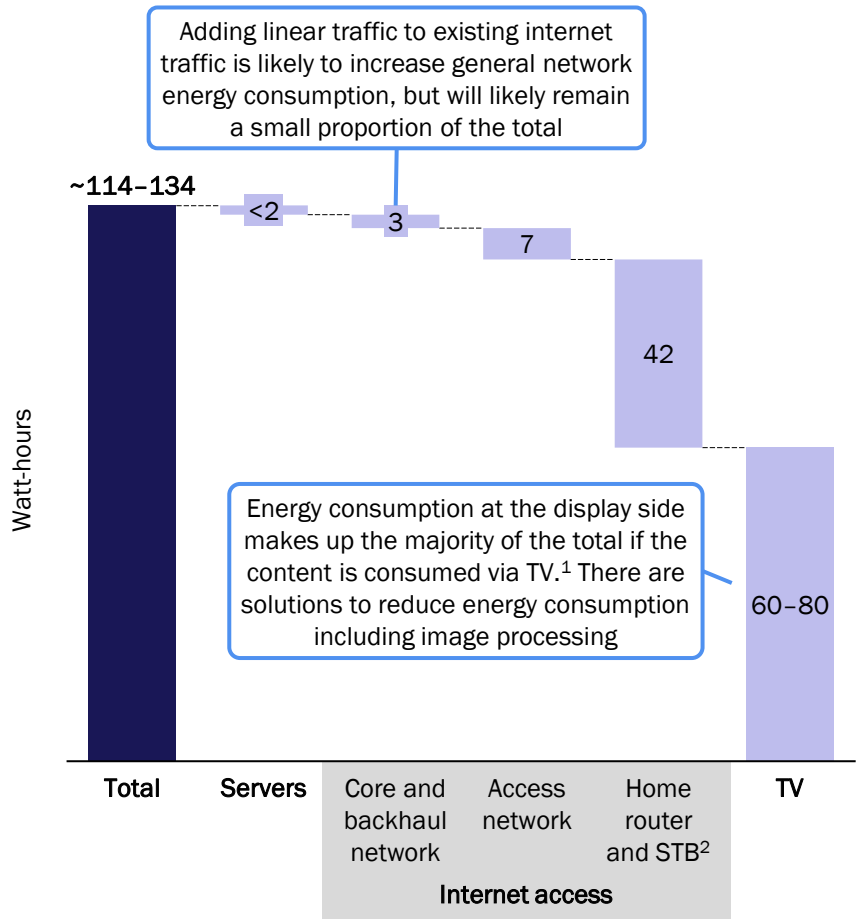
Sharing of server equipment between content providers (e.g. between PSBs) helps to reduce the maximum peak-to-average ratio (i.e. by raising the average, and increasing equipment utilisation)

Sharing of server equipment between CDNs is a possibility, by creating virtualised server functionality. There is a trade-off between the utilisation gains, and virtualisation hardware being less efficient per bit than custom-built equipment

CDN providers have existing commercial models to manage peak loads on their equipment, including caps and/or additional charges for peak-to-average events above a certain ratio (e.g. 4:1)

- These techniques should be considered in the context of streaming servers contributing relatively little to the total energy consumption of an IP-based linear content service (see right)

Energy consumption per hour of IP video streaming for a single viewer in one household

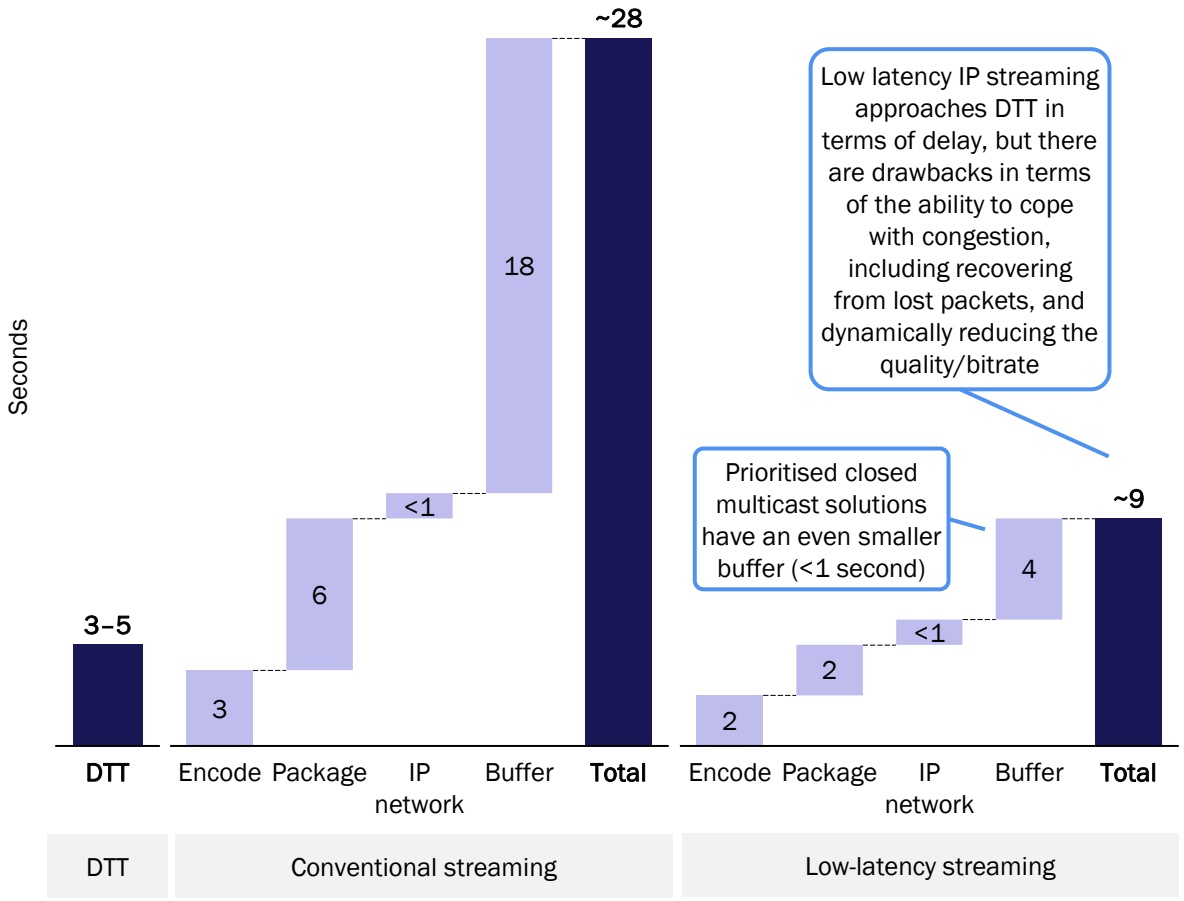


[1] Energy consumption for viewing via a laptop is similar to that via TV at around 70 watts; energy consumption of other displays is much lower: ~1 watts for smartphone and ~3 watts for tablet [2] STB = set-top box

Source: Carnstone, Analysys Mason

Low-latency streaming technologies are emerging but there would be trade-offs between latency and quality and continuity of streaming experience

Video latency comparison between broadcasting and streaming



- Linear streaming over the internet currently incurs a latency of 30 seconds or more: significantly higher than DTT
- The key drivers are packaging and buffering, which are linked
- With conventional streaming, once a segment size has been defined by the packager, it will typically be multiplied by three at the other end for the buffer, which helps to accommodate congested and variable network conditions
- With low-latency streaming, segment size[1] and the number of segments in the buffer are reduced
- These changes reduce latency, but create a trade-off:
  - lower ability to recover from lost packets
  - less time for ABR[2] to identify lower bitrates and switch to a lower quality
- As networks become more bandwidth capable and reliable, low-latency streaming will become more feasible to implement

[1] Streaming Video Alliance (SVA) notes that the trade-offs between latency and quality are best optimised at the two-second segment size | [1] ABR = Adaptive Bit Rate

Source: SVA, Analysys Mason

Legal notice

Contact details

Andrew Daly

Principal

andrew.daly@analysysmason.com

Bonn
Tel: +49 176 1154 2109
bonn@analysysmason.com

Cambridge
Tel: +44 (0)1223 460600
cambridge@analysysmason.com

Dubai
Tel: +971 (0)4 446 7473
dubai@analysysmason.com

Dublin
Tel: +353 (0)1 602 4755
dublin@analysysmason.com

Hong Kong
hongkong@analysysmason.com

Kolkata
Tel: +91 33 4084 5700
kolkata@analysysmason.com

London
Tel: +44 (0)20 7395 9000
london@analysysmason.com

Lund
Tel: +46 8 587 120 00
lund@analysysmason.com

Madrid
Tel: +34 91 399 5016
madrid@analysysmason.com

Manchester
Tel: +44 (0)161 877 7808
manchester@analysysmason.com

Milan
Tel: +39 02 76 31 88 34
milan@analysysmason.com

New Delhi
Tel: +91 124 4501860
newdelhi@analysysmason.com

New York
Tel: +1 212 944 5100
newyork@analysysmason.com

Oslo
Tel: +47 905 59 075
oslo@analysysmason.com

Paris
Tel: +33 (0)1 72 71 96 96
paris@analysysmason.com

Singapore
Tel: +65 6493 6038
singapore@analysysmason.com

Stockholm
Tel: +46 8 587 120 00
stockholm@analysysmason.com


linkedin.com/company/analysys-mason


@AnalysysMason

youtube.com/AnalysysMason