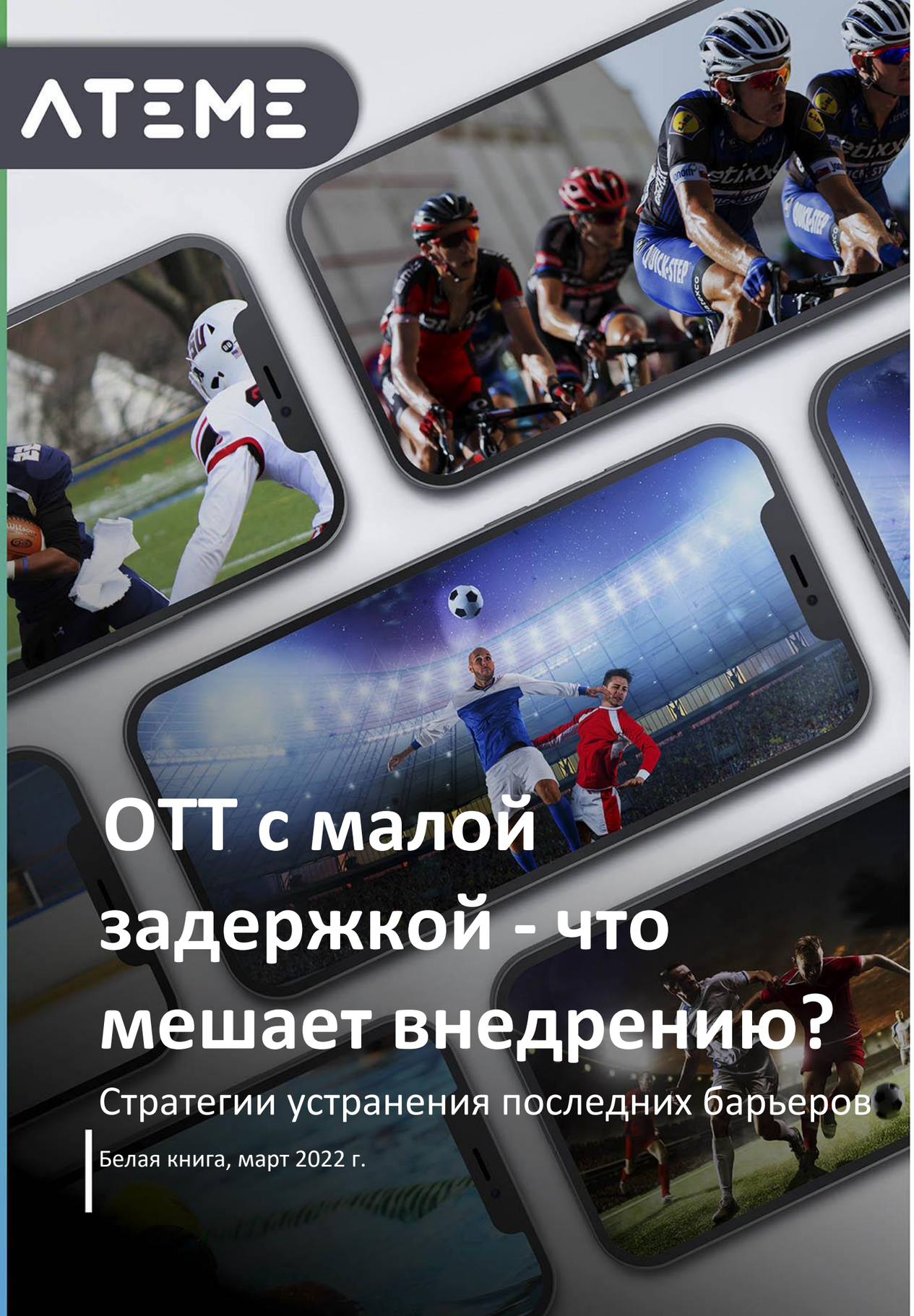


OTT с малой задержкой - что мешает внедрению?

Стратегии устранения последних барьеров

Белая книга, март 2022 г.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Вы смотрите финальный футбольный матч в прямом эфире на большом экране, а ваши соседи радуются забитому мячу за 40 секунд до того, как вы это увидите. Это классическая, граничащая с клише ситуация, которую обычно придумывают, чтобы проиллюстрировать, как снижается качество восприятия при просмотре OTT-контента в реальном времени.

Вопрос в том, почему эта проблема до сих пор не решена?

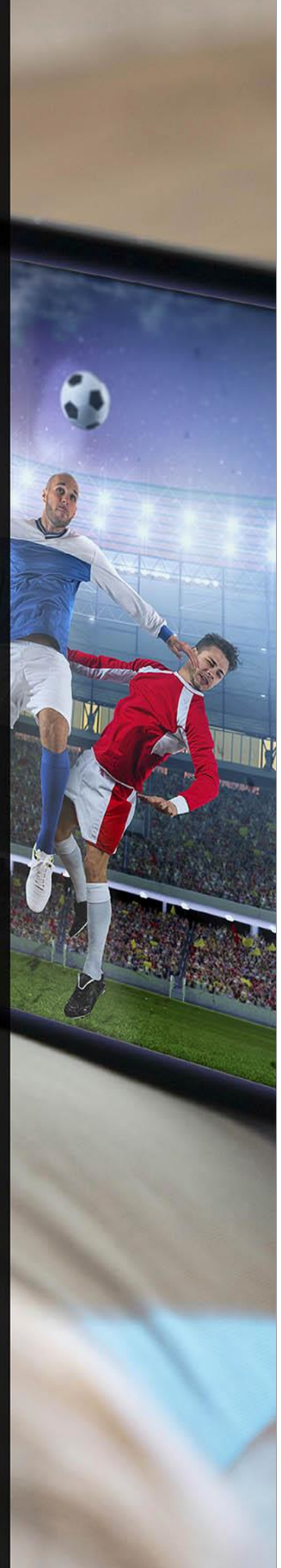
Это большая загадка, если учесть, что существуют технологии для доставки потокового контента в реальном времени в масштабе и с задержкой на уровне традиционной трансляции.

Почему это все еще происходит сегодня? За последние несколько лет многие технологические компании объявили о решениях для потоковой передачи OTT с малой задержкой, но зрители по-прежнему сталкиваются с прежней проблемой.

Несмотря на растущую популярность потоковых сервисов, сокращение задержки OTT остается проблемой как для поставщиков контента, вещательных компаний и поставщиков услуг, так и основным источником разочарования для аудитории. Есть ли способ эффективно снизить задержку в больших масштабах при сохранении качества изображения? Реально ли сегодня добиться низкой задержки в OTT вещании? Ответ – решительное да! OTT вещание с малой задержкой не только теоретически - это реально работает!

В этом техническом документе мы рассмотрим, что блокирует низкую задержку, и как, наконец, устранить эти барьеры. Мы также рассмотрим дальнейшие улучшения, которые, как ожидается, сделают развертывание вещательной системы с низкой задержкой еще проще и повысят ее ценность.

Мы покажем, как вы можете достичь обещанной пятисекундной задержки, полагаясь на сквозную доставку видео OTT от Ateme — уже сегодня.



В чем причина возникновения задержки и что это такое?

Прежде чем мы начнем, давайте определим понятие задержки. Задержка — это время, необходимое видеоконтенту для перемещения от камеры до вашего экрана. Эта задержка возникает на различных этапах, необходимых для обработки и доставки видеоконтента в «потоковом конвейере». Некоторые компоненты этой цепочки уже работают очень близко к своему физическому пределу, то есть к скорости света, с задержками обработки менее миллисекунды на передающей и приёмной стороне.

Другие компоненты, такие как видеокодер, выпускаются в различных модификациях, многие из которых поддерживают режим кодирования с малой задержкой, чтобы уменьшить задержку между приемом последнего входного пикселя и началом передачи первого закодированным байта.

Типовая задержка для линейного вещания составляет от пяти до десяти секунд, тогда как задержка при мультиэкранном OTT вещании исторически составляла от 30 до более 60 секунд, в зависимости от вида устройства для просмотра и используемого рабочего процесса. Вот почему, когда вы смотрите футбол на «втором экране» (или на телевизоре, подключенном к «основному экрану»), вы видите гол на 30-40 секунд позже, чем ваш сосед, который смотрит ту же игру на обычном телевизоре с классическим вещанием.

Задача отрасли состоит в том, чтобы уменьшить эту задержку до значения, близкого к задержке линейного вещания. В этом документе мы рассмотрим различные примеры, для которых задержка будет примерно соответствовать задержке при прямой трансляции, другими словами, будет лежать в 5-секундном диапазоне.

Задержка является критической проблемой для зрителя с точки зрения качества восприятия, но достижение очень низкой задержки за счет плохого качества изображения также не повысит качество восприятия. Концентрируя своё внимание на малой задержке по всей цепочке доставки контента, очень важно не потерять из виду картину в целом и не допускать снижения качества изображения.

ЧАСТЬ 1

Что мешает низкой задержке и как устранить барьеры

A: Развитие форматов

Двумя наиболее широко используемыми форматами потоковой передачи мультимедиа являются Dynamic Adaptive Streaming over HTTP (DASH) и HTTP Live Streaming (HLS). DASH разработан MPEG, а HLS в основном используется Apple для своей экосистемы. Хотя оба доставляют данные схожим образом, они несовместимы друг с другом. Чтобы доставлять одни и те же исходные аудио- и видеоданные в обоих форматах, необходимо дважды сохранить их на кэш-серверах сети доставки контента (CDN) и дважды передать их в несколько разных представлениях.

Формат Common Media Application Format (CMAF), стандартизированный в 2018 году MPEG и принятый в 2016 году Microsoft и Apple, решил эту проблему, сделав MPEG-DASH работоспособным и совместимым с HLS. Упакованные в контейнер CMAF закодированные данные необходимо сохранить и передать в потоковом режиме только один раз. Это уменьшает объем хранилища и снижает требования к пропускной способности, а это означает, что единственная разница между HLS и DASH — это манифест (формат файла), а сам носитель остается прежним.

Низкая задержка и CMAF

При потоковой передаче в любом формате медиафайлы передаются сегментами, каждый из которых длится несколько секунд (от двух до шести секунд). Это по своей сути добавляет несколько секунд задержки от передачи до воспроизведения, поскольку сегменты должны быть закодированы, упакованы, доставлены, загружены, буферизованы, а затем воспроизведены плеером. Таким образом, размер сегмента оказывает решающее влияние на минимальную задержку для выполнения всей этой последовательности. Для ускорения доставки видео в рекомендации стандартов CMAF были включены другие функции, прежде всего Chunk Transfer Encoding (CTE), которая является частью спецификаций http/1.1. Это дает возможность делить сегменты на более мелкие фрагменты или «чанки» длительностью несколько сотен миллисекунд (обычно 500 мс). Этот режим с малой задержкой в CMAF позволяет origin/package серверу постепенно создавать и доставлять сегменты, а получателю (плееру) постепенно запрашивать и создавать медиаконтент вместо того, чтобы ждать, пока станет доступен полный сегмент. Каждый фрагмент может быть доставлен через несколько сотен миллисекунд, и проигрыватель может начать отображать контент с буферизацией всего нескольких фрагментов. Такой подход значительно снижает задержку.

CTE впервые был реализован в DASH как DASH с низкой задержкой (LL-DASH) в 2017 году, но реализация в HLS была отложена, что в целом повлияло на замедление принятия рынком протоколов с низкой задержкой. Понятно, что операторы не хотели развертывать решение, которое создавало бы неодинаковые условия просмотра для клиентов, использующих разные устройства. Внедрение версий с малой задержкой только на DASH (охватывающих все устройства Android) означало бы, что клиенты с устройствами iOS (и некоторыми другими устройствами, использующими HLS) были бы лишены возможности работать с малой задержкой.

Apple ответила в 2019 году решением для HLS с малой задержкой, которое, хотя и было обратно совместимо с устройствами, использующими исходный HLS, но требовало применения функций HTTP/2 Push, которые не были стандартными в архитектуре CDN и поэтому не поддерживались большинством из них. Многие поставщики технологических решений для CDN не были готовы переделывать свои системы для поддержки функций Push и заявляли, что LL-HLS не подходит для масштабирования. Немногие решили принять его, и рост протоколов с низкой задержкой снова остановился. К счастью, Apple прислушалась к ним и в начале 2020 года выпустила новую версию спецификаций LL-HLS, которая убрала требование для http/2 Push, сделав его совместимым со стандартной pull архитектурой origin сервера для доставки OTT. Это упростило развертывание технологии за счет облегчения совместимости с LL-DASH. Наконец, рыночные условия созрели для того, чтобы принять и начать масштабировать решения для стриминга с малой задержкой.

B: От различных рабочих процессов - к единому.

Достижение низкой задержки традиционно требовало инвестиций в различные рабочие процессы: один для HLS и один для DASH. Это означало повышенную сложность и, соответственно, большие первоначальные инвестиции.

Внедрение новой версии LL-HLS означает, что весь рынок впервые может воспользоваться преимуществами упаковки Just-in-Time (JIT) как для DASH с малой задержкой, так и для HLS с малой задержкой, а также для версий этих протоколов со стандартной задержкой, что позволяет операторам добиться низкой задержки, используя существующие рабочие процессы.

Что такое JIT Packaging от Ateme

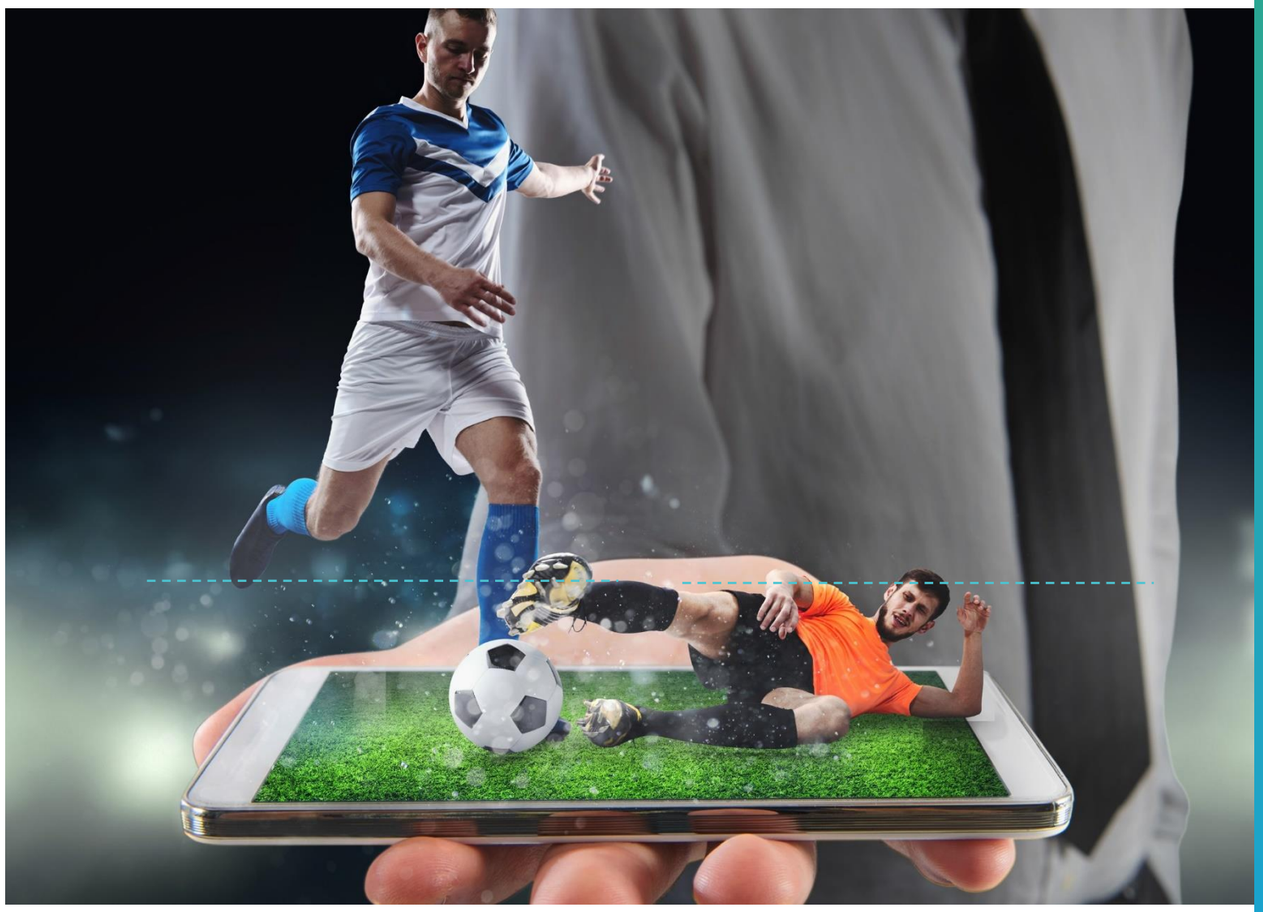
Компания Ateme представила концепцию JIT десять лет назад. Благодаря своей элегантности и эффективности JIT стала стандартной эталонной архитектурой для доставки OTT, поскольку она позволяет значительно сэкономить место на origin сервере, оптимизировать трафик CDN и использовать платформы, ориентированные на будущее развитие.

Packager от Ateme NEA-Live® JIT основан на прорывной концепции, обеспечивающей низкую задержку вещательного уровня в режиме pull как для HLS, так и для DASH. Эта технология позволила компании Ateme первой вывести на рынок JIT-упаковщик с малой задержкой и, тем самым, решить сложную техническую проблему, которую Ateme преодолела, используя собственную запатентованную технологию. Последние усовершенствования NEA-Live имеют огромное значение для поставщиков услуг потоковой передачи. Теперь они могут использовать упаковщик для доставки контента в реальном времени в режиме с малой задержкой, а также пользоваться большей эффективностью JIT-упаковки в режиме pull, когда создается и доставляется только формат, необходимый для запрашивающего устройства. Кроме того, они могут использовать тот же рабочий процесс для других услуг OTT, включая вещание со сдвигом во времени, возврат на начало программы (Start over) и видео по запросу на любом устройстве. Таким образом, зрители могут наслаждаться целым рядом полезных и удобных сервисов, в то время как поставщики услуг потоковой передачи OTT получат удовлетворение от более эффективных и менее сложных операций по обработке контента.

Поскольку контент упаковывается только в требуемые форматы, требования как к обработке, так и к хранению снижаются. Это позволяет использовать меньшее количество серверов и понизить энергопотребление, и, как результат, обеспечивает снижение эксплуатационных расходов и уменьшение воздействия на окружающую среду.

Раньше операторы неохотно вкладывали средства в технологии, которые обслуживали бы только половину рынка, теперь с пакетайзером JIT с малой задержкой от Ateme поддержка всего спектра решений стала намного проще и экономичнее, поскольку для достижения низкой задержки можно использовать традиционные рабочие процессы.

Один из крупнейших клиентов Ateme в Европе, Canal+, недавно воспользовался преимуществами своей существующей платформы NEA-Live, чтобы обеспечить пакетирование с малой задержкой, сделав ее доступной для своих подписчиков. В результате подписчики смогли насладиться малой задержкой при трансляции UHD контента для целого ряда крупных спортивных событий (футбол, Формула-1, регби и т. д.).



C: Важен каждый шаг

Упаковка (Packaging) имеет решающее значение для достижения низкой задержки, но это не единственный элемент в цепочке доставки видео, который может привести к задержке. Для дальнейшей оптимизации необходимо учитывать каждый этап обработки контента: от кодеров до пакетайзеров и CDN. Вот основные этапы рабочего процесса (и стандартные задержки для каждого из этапов):

Кодер - задержка обработки для оптимального качества видео: ~2–5 секунд
Пакетайзер OTT - буферизация для создания сегментов OTT: ~ 2–10 секунд
CDN и origin - ~200 миллисекунд
Плеер - Буферизация при воспроизведении: ~30+ секунд

Анализ показывает, что, хотя кодер и CDN оказывают меньшее влияние, основная задержка возникает из-за буферизации в плеере на стороне зрителя и операций пакетирования HLS или DASH. Для снижения задержки OTT требуется комплексная стратегия, в которой каждый этап играет свою роль.

Кодирование видео

Многие технологии кодирования предлагают различные способы оптимизации задержки. Цель должна состоять в том, чтобы свести к минимуму задержку при максимальном качестве видео даже с контентом 4K HDR. Кодеру необходимо принимать исходные потоки и создавать фрагменты и сегменты нужного размера, которые затем можно загрузить в Origin сервер для доставки.
Ateme сводит задержку к минимуму, сохраняя при этом высокое качество изображения, используя искусственный интеллект для минимизации задержки предпросмотра данных, которая определяет, сколько битов потребуется для кодирования изображения. Кроме того, с помощью искусственного интеллекта мы дополнительно снижаем битрейт, чтобы добиться одинакового качества изображения для OTT и эфирного вещания. Это позволяет нам поддерживать высокое качество сигнала, предлагая все возможности просмотра, обеспечиваемые OTT.

CDN и Origin

Задержка в CDN может быть минимальной, однако вам необходимо убедиться, что CDN поддерживает режимы LL-HLS или LL-DASH, а в идеале — оба. Оптимизация CDN связана с возможностью выполнять прогрессивную доставку сегментов и, используя такой механизм, как Ateme NEA-Live JIT package, упаковывать данные «на лету», чтобы CDN доставляла поток даже во время его создания.

В этом процессе есть два аспекта:
1. DASH: CDN должна поддерживать кодировку передачи фрагментов (CTE).
2. HLS: CDN должен поддерживать http/2

Низкая задержка будет невозможна, если CDN не поддерживает ни один из этих режимов.

Масштабируемость

Также необходимо, чтобы CDN можно было легко и быстро масштабировать. Дело не только в пропускной способности, но и в наличии функциональности для управления большим объемом одновременных запросов. Это связано с тем, что реализация HLS с малой задержкой увеличит количество HTTP-запросов в CDN. Вместо одного запроса на шестисекундный сегмент у вас есть один запрос на каждый фрагмент, являющийся частью сегмента. Этот фрагмент может занимать всего 500 мс, что означает, что для получения точно такого же контента необходимо 12 HTTP-запросов. Это повлияет на возможность обеспечения низкой задержки из-за увеличения нагрузки на CDN.

Технологии CDN должны быть способны справляться с этой нагрузкой по всей сети, чтобы обрабатывать увеличивающийся поток запросов, иначе низкая задержка будет затруднена. Они должны поддерживать CTE и Http/2.

Масштабируемость также зависит от способности достаточно быстро адаптировать возможности доставки CDN при увеличении трафика. Именно здесь концепция «Elastic CDN», представленная Ateme несколько лет назад, имеет решающее значение: операторы должны иметь возможность быстро увеличивать пропускную способность потоковой передачи во время важных событий, которые генерируют пики трафика. Решение NEA-CDN® от Ateme поддерживает все эти функции на уровне производительности, обеспечивающем масштабное развертывание с малой задержкой, опираясь на полностью виртуализированную программную технологию, обеспечивающую гибкость, необходимую для быстрого масштабирования.

Плеер

Плеер — последний кирпичик в рабочем процессе доставки. Необходимо оптимизировать поведение при запуске и сбалансировать буферизацию и скорость воспроизведения, чтобы загрузка фрагментов и рендеринг всегда происходили как можно ближе к реальному времени. В предыдущих рабочих процессах OTT плееру приходилось буферизовать два-три сегмента, прежде чем он мог начать воспроизведение видео. Если сегменты длились шесть секунд, это влекло за собой задержку не менее 12 секунд, прежде чем видео могло начать воспроизводиться.

Чтобы добиться низкой задержки, важно убедиться, что плеер полностью поддерживает LL-HLS и LL-DASH. Это уменьшит задержку воспроизведения примерно до двух секунд по сравнению с предыдущими 10-15 секундами.

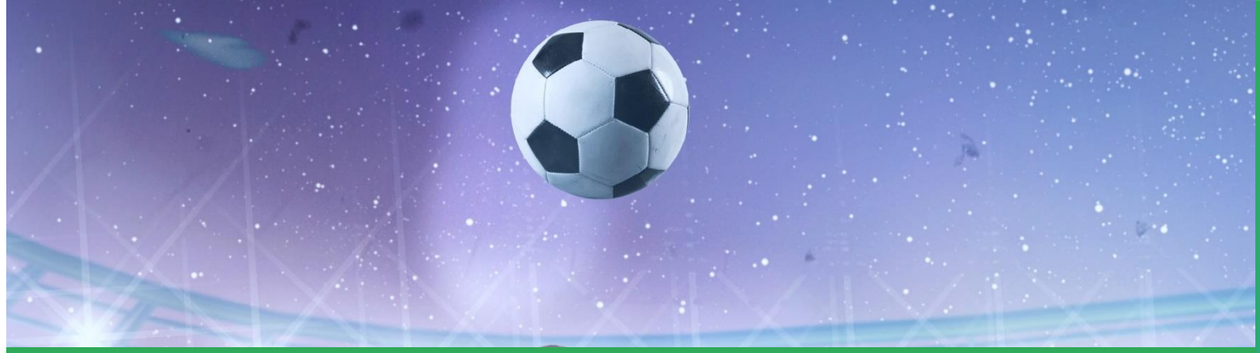
Таким образом, возможность оптимизировать параметры кодирования, уменьшить задержку упаковки, ускорить доставку CDN и уменьшить буферизацию в плеере должна быть применима как к DASH, так и к HLS.

Альтернативные протоколы с низкой задержкой

Важно установить правильные временные рамки для ожидаемой величины задержки. Напоминаем, что здесь мы рассматриваем те случаи, для которых целевая величина задержки находится на одном уровне с задержкой при линейном вещании. Другими словами, она лежит в 5-секундном диапазоне. Для других вариантов использования, например, ставок на спорт, видеоконференций и трансляции непосредственно на стадионе (когда зрители могут посмотреть дополнительный контент и моменты игры на своем устройстве) такая задержка будет непригодна. Для таких применений потребуется задержка около одной секунды или даже меньше и для таких интерактивных приложений существуют альтернативные протоколы.

Однако протоколы, ориентированные на эти приложения, работающие менее секунды, не всегда хорошо масштабируются. Они, как правило, проприетарны и, следовательно, дороги в реализации и лицензировании, а также доступны не на всех устройствах. Например:

- WebRTC не основан на HTTP и не позволяет легко масштабировать сеть (оптимизирован для менее чем 10 000 зрителей) и больше подходит для видеоконференций.
- RTMP - это устаревший протокол, который больше не развивается (например, Adobe больше не поддерживает его).
- SRT ориентирован на рынок сбора и первичного распространения высококачественного контента, но поддерживается небольшим количеством плееров и не предназначен для массовой потоковой передачи OTT миллионам пользователей.

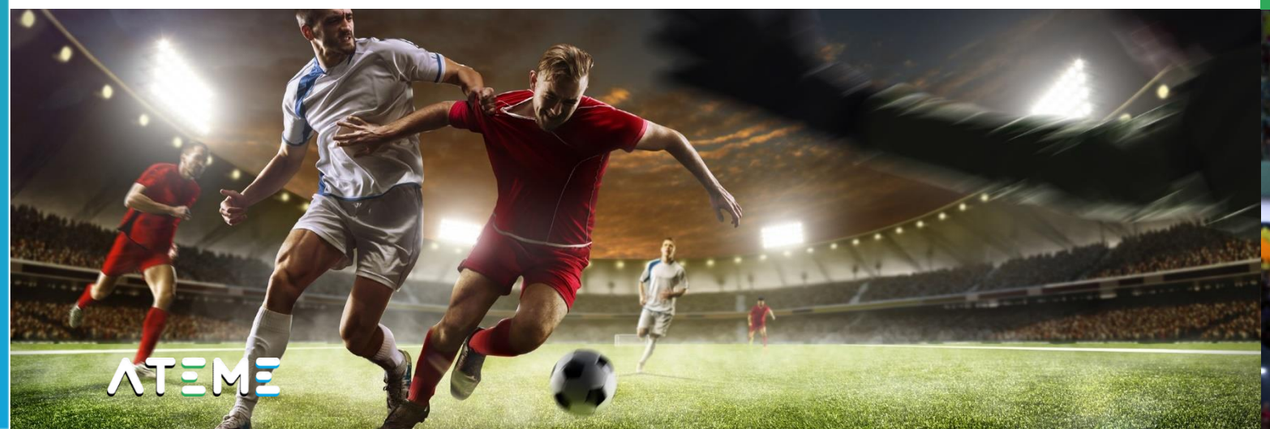


D: Гармонизирующая интеграция

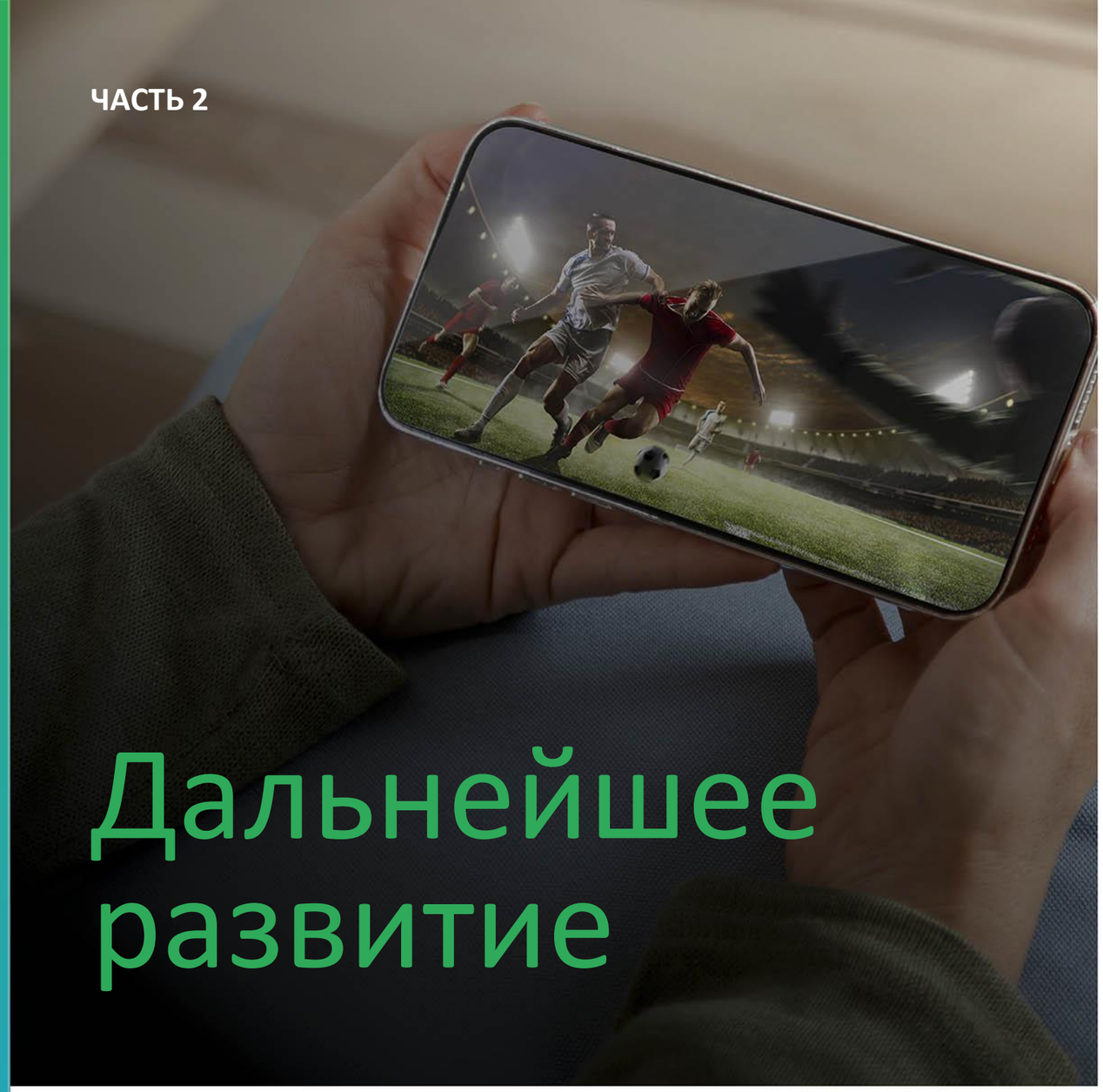
Хотя существуют специальные точечные решения для уменьшения задержки на каждом этапе рабочего процесса, интеграция многих разрозненных компонентов в сквозной процесс может быть сложной, трудоемкой и дорогостоящей. Даже при выборе инструментов и решений, основанных на отраслевых стандартах, при интеграции может потребоваться дополнительная оптимизация и точная настройка, чтобы гарантировать, что все компоненты были тщательно протестированы при совместной работе.

Ateme устранила этот барьер, обеспечив гармоничное взаимодействие своего решения с малой задержкой с экосистемой поставщика. Решение Ateme соответствует отраслевым стандартам и было протестировано и проверено на нескольких плеерах с открытым исходным кодом и коммерческими плеерами для того, чтобы гарантировать, что комплексное решение работает и эффективно снижает задержку. Комплексное решение с малой задержкой, поддерживающее http 1.1 (для CMAF LLC) и http 2 (для LL-HLS) от Ateme включает следующие компоненты:

- Кодер OTT (TITAN) - задержка от 1,5 до 2 секунд
- OTT pull packager (NEA-Live) - задержка от 0,04 до 1 секунды
- CDN (NEA-CDN) - незначительная задержка
- Плеер (сторонний производитель) - ~2 секунды
- Общая задержка от выхода камеры до отображения на дисплее менее пяти секунд



ЧАСТЬ 2



Дальнейшее развитие

Снижение нагрузки на сеть

Как мы видели, внедрение низкой задержки с протоколами DASH и HLS сегодня означает загрузку сети http-запросами, что может негативно сказаться на общей производительности и, следовательно, на качестве просмотра. Если сеть перегружена, она будет неспособна доставлять зрителю видео лучшего качества.

Для решения этой проблемы используется технология получившая название Byterange. Она была разработана для ограничения количества запросов, отправляемых в CDN. Идея состоит в том, чтобы вместо того, чтобы посылать в CDN несколько запросов - для каждого из фрагментов сегмента (например, 12 запросов каждые 500 мс), выдать один запрос для всех фрагментов в начале сегмента, чтобы CDN доставлял все фрагменты по готовности.

Байтовый диапазон может отображаться в плейлисте в виде диапазона байтов 0–1000, за которым следует 1001–2000 и т. д. Заранее определив, какие байтовые диапазоны должны быть доставлены, плеер может отправить только один запрос вместо последовательности запросов. Это ограничивает количество http-запросов, высвобождает часть пропускной способности сети и делает взаимодействие между плеером и CDN значительно более эффективным. Более того, поскольку доставка будет работать одинаково как для HLS, так и для DASH, это позволяет выполнять совместное использование медиафрагментов (чанков) двумя протоколами, что эффективно вдвое снижает требуемую пропускную способность сети.

Технология Byterange может обеспечить как запрос одного сегмента, так и включить передачу DASH с низкой задержкой при кодировании с передачей фрагментов.

Динамическая вставка рекламы

Управление динамической вставкой рекламы (DAI) в потоках с малой задержкой — сложная задача, но многие клиенты заинтересованы в ее решении. Существует неудовлетворенный спрос на решение, которое оптимизирует низкую задержку с DAI, и легко понять, почему. Мероприятия, для которых важна низкая задержка, такие как прямые трансляции спортивных состязаний, привлекают большую аудиторию и, следовательно, являются весьма привлекательным ресурсом для рекламы. Возможность вставлять на лету рекламу с учетом географических и демографических особенностей или персонализированной информации создаст еще больше возможностей для монетизации контента.

У Ateme есть все необходимые компоненты для доставки потоков с малой задержкой и все технологии, необходимые для DAI. Ключевым моментом здесь является Manifest Conditioning — способность Origin packager указывать, где начинается и заканчивается рекламная пауза.

Помечая манифест тегами, мы можем заменить исходный контент в прямом эфире определенными сегментами, ориентированными на конкретного пользователя. Это потенциально создаст один манифест для каждого пользователя (манифест будет указывать на различный рекламный контент для каждого пользователя).

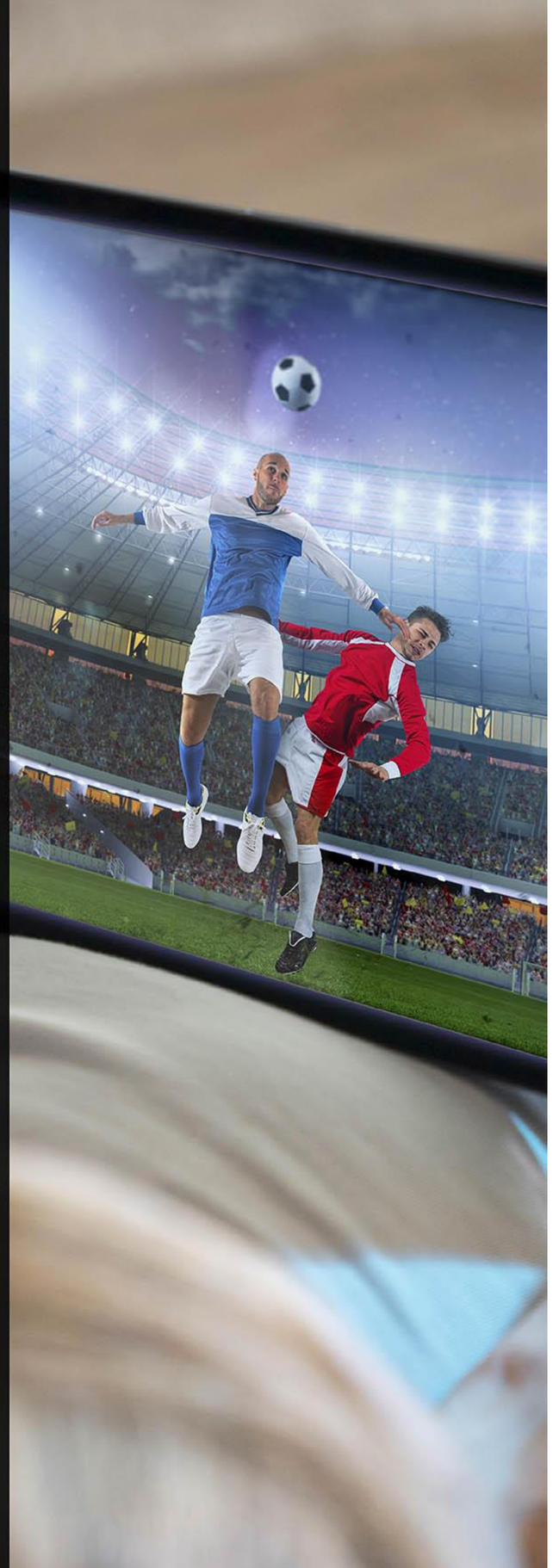
Заключение

Зрители ожидают, что смогут смотреть потоковое видео, в режиме насколько это возможно приближенном к реальному времени, на любом экране. Низкая задержка теоретически возможна уже некоторое время, и многие поставщики уже заявляли об этом. Тем не менее, на практике было развернуто очень мало решений для доставки с малой задержкой.

Основными препятствиями для этого были:

- Низкая задержка была доступна только на устройствах Android, потому что это было возможно только с использованием DASH. Теперь, и с HLS, можно работать со всеми устройствами.
- Стоимость и сложность добавления дополнительного рабочего процесса. Благодаря JIT-упаковщику Ateme с малой задержкой (для HLS и DASH) теперь можно добиться низкой задержки с использованием существующих рабочих процессов.
- Готовность всей цепочки доставки — от кодеров до пакетайзеров и CDN, которая должна поддерживать гораздо большее количество запросов, а также кодирование с передачей фрагментов и http 2.
- Сложность, время и стоимость интеграции различных точечных решений для совместной работы с целью сокращения общего времени задержки. Благодаря комплексному решению Ateme для доставки видео, предварительно интегрированному и протестированному с основными плеерами в экосистеме, низкая задержка теперь является готовым решением.

Параллельно мы также работали над другими решениями, такими как упаковка Byterange, чтобы уменьшить количество запросов, необходимых плееру для воспроизведения контента, и еще больше повысить качество восприятия контента зрителем при минимальной нагрузке на сеть.
Ожидаются дальнейшие усовершенствования, направленные на то, чтобы сделать перспективу низкой задержки еще более привлекательной, монетизировать популярные спортивные мероприятия с помощью целевой рекламы и синхронизировать все устройства для беспрепятственного взаимодействия между экранами.



В результате сегодня вы можете обеспечить требуемую пятисекундную задержку, полагаясь на сквозную доставку OTT видео от Ateme. И хотя некоторые из наших клиентов уже развернули такие службы в LL-DASH, теперь они могут добавить и поддержку HLS с низкой задержкой, чтобы обеспечить идеальное качество для всех зрителей.
Обеспечение прямых трансляций с малой задержкой необходимо для ценного контента, такого как прямые спортивные трансляции. Технология подготовлена и готова к тому, чтобы OTT-провайдеры могли предложить своей аудитории контент такого качества, который они привыкли ожидать от вещательного телевидения.

Authored by Alexandre Arnodin,

Если вы хотите, чтобы ваши зрители наслаждались видеоконтентом с задержкой сопоставимой с линейным вещанием, свяжитесь с нами, чтобы узнать, как мы можем помочь.
Адрес sales@ateme.com

*Дополнительную информацию о развертывании см. в документе “Adoption and Deployment of Internet Streaming Video Technologies”, подготовленном CMAF Industry Forum и поддерживаемом Consumer Technology Association (CTA): https://shop.cta.tech/collections/standards/products/adoption-and-deployment-of-internet-streaming-video-technologies