Облачное гибридное производство и рост расходов на искусственный интеллект

Арсений Ворошилов

Для медиаиндустрии наступили времена, столь же интересные, сколько непростые для вещательных компаний. С одной стороны, они все больше полагаются на производство, в основе которого лежат гибридные рабочие процессы, организованные в частном и/или публичном облаке. С другой же стороны, вещателям приходится сталкиваться с существенными расходами, которые связаны с дефицитом вычислительных ресурсов, обусловленным их интенсивным использованием для нужд искусственного интеллекта. Именно эти расходы выходят на передний план в списке приоритетных задач отраслевых поставщиков.

Новейшие облачные решения, зачастую «заряженные» AI-функционалом, сделали модель гибридного облака как никогда привлекательной на фоне угасающего интереса к использованию только публичного облака или какого-то конкретного специализированного аппаратного решения.

Сложившаяся ситуация требует эффективных подходов, так что мнение представителей лидирующих отраслевых игроков определенно вызывает интерес. К примеру, в Grass Valley полагают, что решение о размещении отдельных компонентов производственных процессов – локально или в облаке – клиенты компании должны принимать с учетом архитектурной сложности каждого из вариантов и на основе оценки целесообразности применительно к решению конкретных задач. В компании уверены, что именно этот подход интересен ее клиентам.

Как и где бы то ни было еще, искусственный интеллект играет все более важную роль и в Grass Valley. Он используется, в частности, для ускорения разработки решений, необходимых клиентам, а также для обеспечения интеграции с системами сторонних производителей, что позволяет клиентам добавлять свои собственные слои AI-приложений.

Необходимость обеспечения безопасности облачных операций, включая анализ уязвимостей, подобный тому, который выполняет новая модель Anthropic Mythos AI, также увеличила потребность в вычислительных ресурсах для искусственного интеллекта. В Grass Valley готовы сделать эту технологию доступной, как только Anthropic снимет действующие ограничения на ее использование. Если вообще снимет, разумеется.

Нельзя упускать из внимания и расходы, связанные с работой уже обученной модели искусственного интеллекта или нейросети с применением новых, ранее не используемых данных для получения результата. Этот процесс получил название инференс (inference). В конце прошлого года авторитетные источники подчеркнули масштаб проблем, с которыми сталкиваются все отрасли. А проблемы эти связаны с использованием искусственного интеллекта. В частности, отмечается, что предприятия практически во всех секторах обнаруживают, что имеющиеся у них инфраструктуры концептуально не соответствуют специфике AI.

Это несоответствие обусловлено инференсом – нескончаемой обработкой, которая необходима по мере того, как AI-модели генерируют новую информацию на основе меняющихся условий. И хотя стоимость инференса за последние два года сократилась в 280 раз, стремительный рост общих расходов на искусственный интеллект оказался еще больше, что легло дополнительным бременем на пользователей. Как результат, зависимость от высокопроизводительных вычислений в публичном облаке может стать слишком дорогостоящей.

Такого мнения придерживаются, например, в компании Ross Video. Там же уверены, что игроки рынка очень быстро адаптируются к новым реалиям. Но в настоящее время для извлечения выгоды из использования AI требуется подключение среды работы с медиаданными к публичному облаку, где обработка буквально пожирает все ресурсы GPU, из-за чего цены непрерывно растут.

Вклад Ross Video в развитие облачных и AI-технологий состоит в том числе в поддержке облачного создания и мониторинга графики XPression в режиме реального времени через масштабируемый сервис по запросу XPression.cloud. Он может использоваться как альтернатива локальному оборудованию XPression. В XPression.cloud искусственный интеллект применяется для идентификации оптимальных шаблонов XPression и для ускорения поиска графических ресурсов, их создания и генерации. В компании уверены, что XPression.cloud обес-

печивает пользователям более эффективный способ управления графикой собственными силами без применения дополнительных средств.

Не отстает от других лидеров и Matrox Video в движении по направлению к гибридным конвергентным решениям, оптимизированным для работы в облаке. В течение последних двух лет компания продвигает собственную медиа-ориентированную транспортную платформу ORIGIN как решение для управления высокопроизводительной многопротокольной передачей данных и масштабированием без использования специализированных аппаратных средств.

До недавнего времени не наблюдалось большого количества реальных сценариев использования этой технологии, но в нынешнем году, отмечают в компании, к числу пользователей присоединились такие поставщики отраслевых решений, как Amagi, EVS и TSL, а также канадский мультисервисный гигант Rogers Communications. ORIGIN выгодно отличается от многих других решений, потому что Matrox успешно использует AI, чтобы упростить клиентам внедрение платформы в свои решения и сервисы. Фактически AI здесь применяется не как часть ПО, а в виде отдельных инструментов, повышающих эффективность работы на каждом конкретном проекте.

Такие варианты использования AI на этапе разработки новых решений поставщиками ускоряют создание систем – как облачных, так и аппаратных. И хотя подобные приложения требуют постоянных затрат на инференс, свойственных ИИ-системам, задействованным в рабочих процессах прямых трансляций, они зависят от ситуации с вычислительными мощностями и от стоимости оборудования. Такого мнения придерживаются в Appear.

Как поставщик оборудования компания придает большое значение цепочке поставок в сфере AI. Здесь искусственный интеллект применяется не только для разработки и создания продукции, но и для поиска новых возможностей и подготовки коммерческих предложений. В частности, для решения проблемы с дефицитом микросхем памяти, с которой сталкиваются практически все производители оборудования и систем для медиаиндустрии.

Appear также вышла на рынок гибридного облачного производства, недавно представив шлюз VX Media (VXM) Gateway. В нем применены интерфейсы с открытым исходным кодом для обеспечения совместимости с решениями других производителей и поддержки стандартной облачной инфраструктуры. Это расширяет возможности распределенного REMI-производства, обеспечиваемого аппаратной платформой X Production. Однако подчеркивается, что VXM не заменяет XP, а служит ее расширением. Выпуск компанией VXM, даже на фоне очередного крупного успеха в области аппаратных решений – Fox Sports использовала платформу X для проведения своих REMI-трансляций Чемпионата мира по футболу 2026, отражает растущий интерес к возможностям виртуализированного гибридного облака во всей традиционной цепочке поставок оборудования для телевизионного производства.

Еще одним ярким примером служит расширенная поддержка облачного производства, воплощенная в виде устройства Nevion компании Sony Electronics. В устройство интегрирована MOXELA – новая облачная программная платформа, содержащая облачный уровень администрирования медиаресурсов VideoIPath, который обеспечивает передачу, обработку и мониторинг производственных процессов с охватом сигналов от камер и другого оборудования, поставляемого Sony и другими производителями.

Еще одна примета времени – это рост спроса на решение нового поколения для потокового вещания видео. От такого решения ждут способности преодолеть ограничения по задержке и однонаправленности, присущие доминирующей сегодня архитектуре потоковой передачи. В связи с этим усиливается поддержка нового стандарта под названием MOQ (Media over QUIC).

Рабочая группа Internet Engineering Task Force еще находится в процессе завершения разработки спецификаций MOQ. Но уже есть несколько проектов, в рамках которых тестируется этот новый подход потоковой доставки множеству потребителей с разными значениями задержки. Интерес и к стандарту, и к результатам испытаний высок.

Эта тенденция ярко проявилась на апрельской выставке NAB Show, где состоялись многочисленные демонстрации MOQ, а сессия конференции, посвященная этому стандарту, прошла с полным аншлагом.

В число компаний, анонсировавших поддержку MOQ на уровне инфраструктуры, вошли операторы публичных сетей доставки контента Broadpeak, CacheFly, Cloudflare, CDN77 и Vindral, а также поставщики облачных вычислительных платформ Akamai, Amazon Web Services (AWS) и Oracle Cloud Infrastructure (OCI).

Не отстают и поставщики программных решений для стриминга, продемонстрировавшие передачу медиаданных и расширенные функции через инфраструктуру с поддержкой MOQ, такие как Ant Media, Ateme, Bitmovin, EZDRM, Nomad Media, Norsk Media, Oracle Media Services, Synamedia, Red5 и Wowza.

Хотя специализированные системы потоковой передачи в режиме реального времени, поддерживающие видеоконференции и множество других приложений, существуют уже несколько лет, MOQ призван удовлетворить потребность в стандартизированной и повсеместно доступной двунаправленной «супермагистрали», по которой видеоданные могут передаваться в любом направлении и в любом масштабе, с задержкой, адаптированной к требованиям конкретного сценария использования. Именно поэтому Red5 – разработчик высокомасштабируемой стриминговой платформы на базе WebRTC – активно поддерживает MOQ.

WebRTC изначально разрабатывался как одноранговый (peer-to-peer) протокол, а это означает, что для масштабирования отсутствует стандартный подход, подобный тому, который используется при построении кластеризованных облачных ресурсов. MOQ изначально создавался как масштабируемая технология вещания для Интернета, что позволит ускорить внедрение потоковой передачи нового поколения.

Базовый набор спецификаций MOQ, известный как MOQ Transport (MOQT), представляет собой несовместимый с традиционными подходами к потоковой передаче

принципиально новый вариант. В частности, такие технологии, как HLS и MPEG-DASH, основаны на протоколе HTTP (Hypertext Transfer Protocol). MOQT, будучи так называемой безустановочной (connectionless) платформой, опирается на простой подход publish/subscribe (публикация/подписка) к потоковой передаче: он сигнализирует о начале и завершении сессии, избегая тысяч обменов запросами и ответами между клиентами и серверами, которые выполняются при потоковой передаче по HTTP.

Новый стандарт также позволяет автоматически настраивать сквозную задержку до уровня, необходимого для конкретного применения – в диапазоне от 200...400 мс в режиме реального времени, что практически незаметно для человека, до того, что называют «интерактивным live», то есть задержки 2 с, подходящей для спортивного тотализатора, онлайн-казино, банковских и аналогичных приложений. Предусмотрен также режим conservative live с задержкой 5 с, обеспечивающий стабильное качество HD, 4K и в перспективе 8K.

Параллельно с разработкой MOQT силами IETF и тесно связанной с ним неформальной организации OpenMOQ Software Consortium реализуется ряд других инициатив, направленных на создание практической программной среды для работы транспортного протокола. В их число входят разрабатываемые в IETF форматы потоковой передачи на уровне медиаслоя, новый шаблон клиентского плеера, созданный Red5 и поддержанный консорциумом, а также несколько готовых к развертыванию плееров.

Основная проблема, которую решают с помощью MOXELA, заключается в потребности производителей получить гибкость в адаптации к различным рабочим процессам, запуская различные функции обработки для разных потоков параллельно в локальных и общедоступных облачных кластерах. Так считают специалисты Sony Electronics, специализирующиеся на оборудовании и решениях Nevion.

MOXELA совместима с распространенными стандартами, такими как SMPTE ST 2110 и NDI, а также с формирующимися отраслевыми рекомендациями, разрабатываемыми в рамках инициативы Digital Media Facility (DMF) Европейского вещательного союза (EBU). В их числе – Media eXchange Layer (MXL), предусматривающий поддержку контейнеризированных микросервисов для обмена несжатыми аудио, видео и метаданными в режиме реального времени и непосредственно в памяти между программными приложениями различных производителей. Эксперты рассматривают это как большой шаг на пути к облачному проведению спортивных трансляций и производству других программ, однако впереди еще много работы.

В настоящее время идет переход к облачному производству от уровней Tier 3 и Tier 2 к Tier 1, но для завершения этого перехода все еще требуются определенные технические разработки, пока еще не доведенные до конца. В целом же речь идет о поэтапной дорожной карте, которая также предусматривает внедрение AI для автоматизации решения сложных задач по поиску и устранению неисправностей, планированию и настройке, связанных с крупными облачными производственными системами.

Реализация MXL в линейках продукции различных производителей, чему способствует выпуск Linux Foundation первого комплекта средств разработки MXL (SDK), быстро становится отраслевой нормой. Потому что переход к гибридному облачному производству сделал оперативную активацию стандартных компьютерных средств (COTS) с обеспечением совместимости между изделиями разных производителей необходимым условием для сдерживания затрат.

Sony, а также недавно объявившая о поддержке протокола компания Ross Video, дополняют список сторонников MXL, в который входят Appear, AWS, Grass Valley, Imagine Communications, Intel, Lawo, Matrox, NVIDIA, Riedel Communications и Telos Alliance.

Для Lawo растущее отраслевое признание гибридного подхода к вещательному производству полностью соответствует многолетнему продвижению компанией платформы управления и администрирования Lawo HOME. Платформа работает на стандартных серверных системах в частных или публичных облаках и обеспечивает обнаружение, аутентификацию, управление безопасностью и мониторинг всех подключенных к производству устройств через единый пользовательский интерфейс.

Чтобы снизить неопределенность в отношении затрат по мере распространения модели гибридного облачного производства, Lawo представила Flex – систему оплаты на основе кредитов, которая позволяет клиентам легче адаптироваться к изменяющемуся балансу используемых ресурсов в частном и публичном облаках.

Кредиты Flex пользователь может покупать и использовать как фишки в казино, тратя их на любое приложение HOME. Гибридная облачная модель становится все более привлекательной, но сохраняющаяся неопределенность требует предоставлять клиентам как можно больше гибкости в выборе способов расходования средств на облачные ресурсы.

Например, клиенты предпочли бы максимально использовать собственные ресурсы, однако в настоящее время серверы почти втрое подорожали из-за дефицита микросхем, а срок их поставки может достигать полугода. Параллельно в публичных облаках наблюдается рекордный рост стоимости обработки данных – к третьему кварталу рост уже составляет 80%.

Если учесть дополнительные сложности, вызванные растущим спросом на ресурсы AI, становится очевидно, что гибридная облачная модель может оказаться не таким безусловным способом экономии, каким представлялась многим прежде. Однако в условиях производства контента в прямом эфире, где необходимо поддерживать разрозненные рабочие процессы на базе ST 2110, NDI, SDI и других трактов, а также обеспечивать работу через браузер для решения задач монтажа из разных географических точек, преимущества облачной доступности, гибкости и эффективности слишком значительны, чтобы ими пренебрегать.

Это повышает значимость компетенций в области оптимизации затрат на облачные ресурсы. Как отмечают эксперты, в условиях растущей неопределенности среди производителей контента понимание того, какое решение лучше всего подойдет конкретному заказчику, во многом позволяет ставить правильные вопросы.