Cloud Advisor

Оптимизация расходов на облачную инфраструктуру

Автор: Cloud Advisor

Тип документа: Whitepaper

Содержание

Введение

По данным исследовательской компании IDC, во втором квартале 2020 года расходы на ИТ-инфраструктуру публичных облаков впервые превысили расходы на традиционную ИТ-инфраструктуру. Эксперты IDC считают, что и в дальнейшем затраты на облачные ресурсы будут продолжать расти. Все больше организаций перемещают свои информационные и вычислительные мощности в облако и вопрос экономии на стоимости облачных услуг становится все более актуальным.

> Респонденты оценивают потери собственных вложений в облако на уровне 30%. **Однако, с высокой долей вероятности можно утверждать, что фактический процент даже выше,** так как многие организации склонны недооценивать объемы потраченных впустую средств.
>
> Отчет State of the Cloud Report 2022, Flexera

В связи с особенностями модели предоставления вычислительных ресурсов в облаке, а также тем фактом, что методы планирования в облаке достаточно сильно отличаются от традиционных, далеко не всегда компании способны правильно спланировать инфраструктуру. Это приводит к неэффективному использованию мощностей, и, соответственно, завышенным, неоптимальным расходам пользователей.

По разным оценкам, от 30% до 70% всех расходов пользователей в публичных облаках тратятся впустую. Проблема оптимизации затрат на облачную инфраструктуру давно является одной из наиболее важных для пользователей.

К счастью, существует достаточное количество практик для снижения и оптимизации затрат в облаке и ниже мы перечислим ряд наиболее эффективных из них, а также коснемся особенностей их применения. Нужно учитывать, что не все методы универсальны – некоторые из них могут быть использованы в зависимости от задач и профилей потребления конкретных проектов.

Найдите неиспользуемые или «забытые» ресурсы

Наиболее очевидным и простым способом оптимизации является обнаружение и высвобождение неиспользуемых ресурсов. Удостоверьтесь, что в вашей инфраструктуре не затерялись ресурсы, оставленные без присмотра – например, после завершения какого-нибудь проекта, или просто не удаленные вовремя виртуальные машины, диски, инстансы баз данных или любые другие тарифицируемые вашим провайдером ресурсы. Возможно, какие-то из них были созданы автоматически, но не были сконфигурированы на последующее удаление после выполнения назначенных им задач или остались невостребованными после удаления «родительского» ресурса. Такие ресурсы можно удалить или по крайней мере временно остановить, чтобы они не вносили свой вклад в общий счет за облачные услуги вашего провайдера.

По оценкам специалистов Яндекс.Облака, **до 35% расходов** его пользователей на облачные сервисы **можно оптимизировать**.

Оптимизируйте имеющиеся ресурсы

Облачная платформа предоставляет гибкий выбор выделяемых мощностей, таких как количество и доля vCPU, платформа, объем оперативной памяти, размер и тип жесткого диска. Однако, на этапе развертывания инфраструктуры требования к количеству необходимых вычислительных ресурсов могут быть неочевидны. Это приводит к выделению избыточных мощностей для предотвращения проблем с производительностью. Но выбрав изначально некий уровень производительности, вы продолжаете платить за него вне зависимости от того, используете ли полностью выделенные ресурсы, хотя зачастую они используются не на 100%.

По данным Flexera, шестой год подряд **оптимизация расходов является наиболее приоритетным направлением** деятельности пользователей облачных сервисов.

Отчет State of the Cloud Report 2022, Flexera

В дальнейшем, сопоставив изначально выделенные вычислительные мощности и реальные паттерны их использования, можно привести объем ресурсов к реально востребованным значениям, соответствующим нагрузке, и тем самым оптимизировать расходы.

Например, если виртуальная машина использует процессор не более чем на 20% и оперативную память не более чем на 50%, очевидно, что она не получает достаточную нагрузку и, уменьшив объем выделенных ей ресурсов, можно существенно снизить ее стоимость.

Используйте автоматическое масштабирование

Часто специфика проекта подразумевает варьирующуюся нагрузку (простейший пример – зависимость спроса от времени суток или сезона, вызывающая повышение нагрузки при обработке большего числа заказов). При этом нет необходимости постоянно иметь количество ресурсов, достаточное для обслуживания пиковой нагрузки, и, соответственно, платить за них, хотя они не используются существенную часть времени. Отслеживать повышение и понижение нагрузки, чтобы добавить новые или удалить избыточные ресурсы вручную, не всегда представляется возможным/целесообразным.

Облачные платформы предлагают инструменты для автоматического масштабирования на основании заданных вами критериев. Вы просто настраиваете правила или триггеры и сервис будет автоматически включать или выключать виртуальные машины для обеспечения необходимой производительности и предотвращения отказов в обслуживании при скачках нагрузки.

Обязательно воспользуйтесь этими инструментами если тип нагрузки вашего сервиса подразумевает волатильный спрос на вычислительные мощности – это позволит скорректировать их количество в зависимости от текущей нагрузки без вашего вмешательства и существенно сэкономить на покупаемых ресурсах, обеспечив при этом готовность к пиковым нагрузкам.

Отключайте виртуальные машины в периоды простоя

Еще один способ сократить расходы, основываясь на данных о нагрузке на инфраструктуру в различные периоды времени, – временное выключение ВМ если известно, что они не используются в течение определенного промежутка времени. В качестве инструмента анализа для получения информации могут использоваться тепловые карты или другие механизмы мониторинга облачной активности, предоставляемые провайдером.

Тепловая карта наглядно отображает периоды работы и периоды простоя ресурсов, позволяя определить время отключения и запуска ВМ. Например, в случае стандартной рабочей недели, тепловая карта покажет какие серверы, будучи неиспользуемыми в выходные дни и/или в нерабочие часы, могут быть безболезненно отключены в это время. Это может быть сделано как вручную так и автоматизировано с помощью специализированного программного обеспечения, способного реализовать выключение и запуск машин по расписанию. Такой подход позволит значительно сократить время работы ресурсов, и, следовательно, расходы на них.

Инвестируйте в резервируемое потребление

Если ваша инфраструктура обеспечивает прогнозируемые и стабильные нагрузки в течение длительного времени, вы можете воспользоваться возможностью зарезервировать у облачного провайдера определенный объем ресурсов на фиксированный срок (как правило год или три) и получить существенную скидку за обязательство его использовать. Этот способ хорошо подойдет, например, для баз данных и может сэкономить до 50% стоимости ресурсов.

Используйте прерываемые ВМ

Для определенных сугубо вычислительных нагрузок, нетребовательных к отказоустойчивости (таких как обсчет или обработка данных), или кратковременных тестовых задач вы можете использовать прерываемые ВМ. Это свободные вычислительные ресурсы облачной платформы, предоставляемые по меньшей цене при условии, что они могут быть отозваны (принудительно остановлены) в любой момент в течение суток. Остановка может произойти если возникнет нехватка ресурсов для запуска обычных виртуальных машин в той же зоне доступности или по прошествии 24 часов с момента запуска прерываемой ВМ. При этом все данные внутри такой ВМ сохраняются и она может быть запущена заново.

В остальном прерываемые виртуальные машины работают также как обычные и идеально подходят для отказоустойчивых приложений, способных выдерживать возможные прерывания в работе своих экземпляров. Использование такого подхода позволяет сэкономить до 75% стоимости.

Используйте облачные функции

При определенных видах нагрузки вы можете практически полностью отказаться от собственной инфраструктуры в облаке и решать необходимые задачи с помощью т.н. бессерверных вычислений на основе облачных функций. Этот способ подойдет, например, для серверной части мобильных приложений, которые получают данные из БД для последующего преобразования; обработчиков, которые сжимают или иным образом модифицируют объекты при их загрузке в объектное хранилище; интеграции со сторонними сервисами и API или обработки потоковых данных.
При этом вы платите только за то время и тот объем мощности (CPU, RAM), которые необходимы для совершения конкретной операции.

Облачные функции позволяют запускать ваш код в безопасном отказоустойчивом окружении без необходимости создания и обслуживания полноценных виртуальных машин. Расходы за обслуживание и администрирование инфраструктуры в этом случае полностью лежат на провайдере. При этом вы фактически бесплатно получаете автоматическое масштабирование ваших запросов, т.к. при увеличении их числа сервис создаст необходимое количество дополнительных экземпляров вашей функции. Вам не нужно заботиться о планировании и управлении ресурсами, сервис сам выделяет мощности, необходимые для выполнения ваших задач, а вы платите только за время выполнения функции и объем памяти, который она использует.

Храните данные в объектном хранилище

Экономить можно также и на хранении данных. Вместо разворачивания отказоустойчивой системы хранения на основе дублирования ВМ, воспользуйтесь объектным хранилищем. Сервис позволяет оптимизировать расходы на хранение больших объемов данных - например, файлов резервного копирования, видео- и аудиофайлов.

Используйте жизненные циклы объектов

Как правило, объектное хранилище позволяет хранить объекты в разных классах хранилища, имеющих различную стоимость хранения, избыточность и скорость доступа к ним. Хранить все данные в хранилище с высокой степенью доступности совершенно необязательно. Некоторые типы данных меняют паттерны использования в зависимости от стадии их жизненного цикла. Например, какие-то данные могут становиться менее востребованными по прошествии определенного времени и перемещение их в более дешевый класс хранилища позволит существенно сэкономить.

Объектное хранилище позволяет задать конфигурацию жизненных циклов объектов – например, использовать стандартное хранилище для активной работы с объектами первой необходимости, а холодное – для длительного хранения объектов с редкими запросами на чтение. Разработайте стратегию по выбору класса хранилища, соответствующего стадии жизненного цикла ваших данных, и оптимизируйте расходы с помощью автоматического перемещения данных между классами.

Заключение

По опросам пользователей облачных сервисов, оптимизация расходов уже давно является одной из самых важных задач. Облачные провайдеры предоставляют множество инструментов для ее решения, однако разработка эффективной стратегии снижения затрат требует основательного подхода и вовлечения специалистов разного уровня – от DevOps-инженеров и менеджеров продуктов до финансистов и руководства компании.

Во внимание необходимо принять и тот факт, что инфраструктура, решаемые задачи, требования и нагрузка постоянно меняются и задача ответственных – выстроить процесс, направленный на периодическую оценку эффективности затрат, уточнение критериев и поиск новых методов оптимизации. Только в этом случае можно достичь желаемого сокращения расходов.

Отметим, что на практике самостоятельно находить неиспользуемые ресурсы, оптимизировать работающие ВМ и БД, применять политику жизненных циклов в объектном хранилище при существенном размере инфраструктуры просто не представляется возможным. В таких случаях требуется привлечение специализированных решений, способных постоянно сканировать все имеющиеся ресурсы, выявлять неэффективное расходование средств и давать рекомендации по оптимизации затрат.