# эффект загрузки

## экономика облачных вычислений

	Аннотация
В статье раскрывается ав <del>тор</del>	оский подход к
более глубокой экономической суг	щности модели
Облачных Вычислений «pay-as-yo	и-go», которую
можно перевести как принцип пог	пребления услу-
ги в следующей форме – «плата і	исключительно
за использование». При этом учит	нывается, что в
модели Облачных Вычислений пол	ьзователь пла-
тит за фактическое использовани	ие Полезности, а
в модели традиционного датаценг	пра – за исполь-
зование ресурсов оборудования.	

**Ключевые слова:** модель Облачных Вычислений, экономическая эффективность Облачных Вычислений, эластичность Облачных Вычислений, традиционные датацентры

Звестно, что модель Облачных Вычислений существенно более гибкая, чем любая другая модель потребления компьютерных услуг, что означает возможность применения адаптивной стратегии. А именно, для Облачных вычислений возникают довольно широкие возможности исправления ситуации, когда нам нужно оценить риски принятия решений на основе ошибочных прогнозов. Широкие возможности возникают за счёт эластичности модели Облачных вычислений.

Стоимость аппаратных решений постоянно снижается. Однако неравномерно по типам оборудования (например, стоимость центральных процессорных устройста — ЦПУ и устройств хранения снижается быстрее, чем стоимость использования глобальной вычислительной сети — WAN). В модели Облачных Вычислений пользователь платит за фактическое использование Полезности, в модели традиционного датацентра — за использование ресурсов оборудования.

## Макаров С.В.

зам. директора
Центра
«Инновационные
образовательные
технологии»
Высшей школы
корпоративного
управления Академии
народного хозяйства
при Правительстве РФ,
специалист в области
информационных
технологий
sergei.makarov@
gmail.com

в модели Облачных Вычислений пользователь платит за фактическое использование Полезности, в модели традиционного датацентра — за использование ресурсов оборудования

Облачные Вычисления являются своего рода буфером, для пользователей, при изменениях в стоимости оборудования, поэтому можно ожидать, что использование Облака может быть более эффективным, нежели строительство собственного датацентра.

В процессе принятия решения необходимо тщательно оценить ожидаемый уровень средней и пиковой загрузки оборудования. Рассчитать варианты стоимости решений, которые будут использоваться при пиковой нагрузке и какова стоимость эксплуатации этих решений при низкой нагрузке. Рассчитать операционную стоимость, для аналогичных сценариев, для различных технологий Облачных Вычислений.

## Эластичность и управление рисками

Экономическая эффективность Облачных Вычислений, которая «лежит на поверхности» — это «конверсия капитальных затрат в операционные расходы» (СарЕх to ОрЕх). Но более глубокая экономическая сущность модели Облачных вычислений выражается фразой «рау-аs-you-go». По-видимому, можно предложить перевод этой фразы, как принципа потребления услуги, в следующей форме «плата исключительно за использование».

Время использования, купленное в модели Облачных Вычислений, может быть распределено неравномерно (покупка 100 серверо-часов сегодня и ноль завтра — при этом плата за сервер-час не меняется; 100Мb трафика сегодня и ноль завтра — при этом плата за использование пересылки единицы информации не меняется)<sup>1</sup>.

И уже к этому принципу – принципу *pay-as-you-go* добавим *«CapEx to OpEx»*, когда отсутствие необходимости первоначальных вложений на ИТ позволит пропорционально увеличить размеры средств, направляемые на другие нужды.

Таким образом, несмотря на то, что например, цена покупки полезности одного сервера, в течение его

<sup>&</sup>lt;sup>1</sup> См. например: Usage based pricing: Washington Post Case Study: AmazonWebServices. Available from: http://aws.amazon.com/solutions/case-studies/washington-post/.

жизненного цикла, в модели «pay-as-you-go» может оказаться выше, чем стоимость эксплуатации аналогичного сервера в собственном датацентре, мы утверждаем, что экономические преимущества от применения модели Облачных Вычислений существенно выше, за счёт эластичности и возможности управления рисками, в особенности рисками необеспеченных пиковых загрузок и недозагрузки оборудования.

#### Начнём

#### с эластичности

Ключевой фактор Облачных Вычислений – возможность добавлять или удалять ресурс небольшими «порциями». Например:

- а) для низкоуровневой Облачной платформы;
- б) для AWS EC2 один сервер за час (сравните с неделями на покупку и установку нового сервера в датацентре);
- в) для высокоуровневых Облачных платформ один пользователь (плюс один гигабайт для хранения данных) за месяц (сравните с «пожизненной» схемой лицензирования коробочного ПО).

Таким образом, обеспечивается существенно более точное соответствие между нагрузкой (количеством функций за единицу времени) и ИТ-ресурсом, выделяемым для исполнения необходимого количества функций. Для традиционных датацентров утилизация серверов находится в диапазоне от 5% до 20%. Это удивительно мало. Но вполне соответствуют другому статистическому наблюдению, а именно, что пиковая нагрузка превышает среднюю в 2-5 раз.

Средняя нагрузка существенно ниже пиковой, но необходимость обработки ситуации пиковой нагрузки требует наличия соответствующих ресурсов, которые по необходимости будут простаивать в обычные часы. Чем больше разница между пиковой и средней нагрузкой, тем больше бесполезное простаивание ресурсов. Поясним на простом примере, как эластичность Облачных Вычислений позволяет уменьшить бесполезное простаивание, и, таким образом, может компенсировать потенциально большую стоимость часа аренды сервера в сравнении с часом работы собственного сервера (рис. 1).

время использования, купленное в модели Облачных Вычислений, может быть распределено неравномерно...

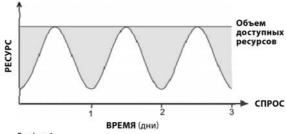


График 1

Рис. 1. Пиковая загрузка обрабатывается.
Но при отсутствии эластичности, мы имеем простаивающие ресурсы (заштриховано), во время непиковой нагрузки<sup>2</sup>

### Пример 1

Предположим, наш сервис имеет вполне предсказуемую дневную нагрузку: 500 серверов для обработки пиковой нагрузки в полдень, и 100 серверов для обработки запросов в полночь (рис. 1) Средняя нагрузка, за день, составит 300 серверов. Утилизация оборудования составит 300×24=7200 серверо-часов. Однако мы должны обрабатывать пиковую нагрузку, поэтому используемые ресурсы, за которые мы должны заплатить, составят 500×24=12000 серверочасов. Разница составит фактор 1,7.

Таким образом, если почасовая аренда сервера за три года (по принципу «pay-as-you-go») будет в 1,7 раза меньше стоимости покупки нового сервера, мы достигнем уровня экономии в модели Облачных Вычислений, в сравнении с традиционной моделью покупки сервера<sup>3</sup>.

Пример весьма прост, а потому не учитывает все преимущества эластичности для реальной жизни. Колебания спроса на достаточно сложные сервисы в реальной жизни могут быть как краткосрочными (как в примере) или долгосрочными (месячными, сезонными, годовыми), так и непредсказуемыми (реакция пользователей на внешние, «новостные» факторы).

чем больше разница между пиковой и средней нагрузкой, тем больше бесполезное простаивание ресурсов

<sup>&</sup>lt;sup>2</sup> Иллюстрация http://www.eecs.berkeley.edu/Pubs/Tech Rpts/2009/EECS-2009-28.html

<sup>&</sup>lt;sup>3</sup> Пример расчётов смотри «Amazon.com CEO Jeff Bezos on Animoto» http://blog.animoto.com/2008/04/21/amazon-ceo-jeff-bezos-on-animoto/.

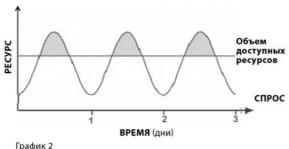


График 2

*Рис. 2.* Пиковая загрузка не обрабатывается, случай 1. Необработанные запросы (клиенты) приводят к недополученной прибыли<sup>4</sup>

В традиционной модели приобретение и наладка сервера для обработки нового уровня загрузки занимает недели, поэтому единственный способ - это планировать ввод оборудования в эксплуатацию заранее. Даже при правильном прогнозировании и планировании, как в приведенном примере, это может привести к существенной недозагрузке оборудования в обычные часы (дни, месяцы). При неправильном прогнозировании, при недооценке загрузки, ситуация становится ещё хуже (*puc. 2*).

При недооценке загрузки мы попросту не обслуживаем поступающие запросы возможных клиентов нашего сервиса. В случае переоценки количества запросов пиковой загрузки, достаточно легко посчитать потери от простоя оборудования. В случае недооценки загрузки, посчитать потери несколько сложнее. Кроме эффекта недополученной прибыли от неоказания сервиса потенциальным клиентам, возникает долговременный эффект от невозврата потенциального клиента, который был неудовлетворён низким качеством сервиса (или вообще, его неоказанием).

Эта ситуация проиллюстрирована на рис. 3, когда количество запросов (клиентов) уменьшается, пока не достигнет уровня, когда мы не начнём удовлетворять запросы всех пользователей. Однако наблюдается тенденция уменьшения количества клиен-

в традиционной модели приобретение и наладка сервера для обработки нового уровня загрузки занимает недели

<sup>&</sup>lt;sup>4</sup> Иллюстрация http://www.eecs.berkeley.edu/Pubs/Tech Rpts/2009/EECS-2009-28.html

качество эластичности весьма ценно не только для стартапов, но и для крупных компаний

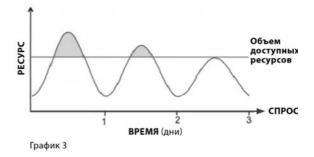


Рис. 3. Пиковая загрузка не обрабатывается, случай 2. Неудовлетворённые клиенты уходят из сервиса<sup>5</sup>

тов, что в долгосрочном масштабе может принести большие неприятности.

# Пример 2

Используем, в качестве примера, тот же случай с компанией Animoto, который мы уже рассмотрели. Animoto запустил свой сервис в Facebook . Спрос на сервис вырос за три дня в размерах, потребовавших увеличить количество серверов с 50 до 3500. Предсказать удвоение требуемых ресурсов каждые 12 часов в течение трёх дней было практически невозможно. Через некоторое количество времени спрос упал существенно ниже уровня пиковой нагрузки.

В данном случае, вопрос эластичности – масштабируемости вверх – не лежал в сфере оптимизации стоимости использования оборудования, он лежал в области обеспечения функционирования сервиса вообще. А масштабируемость вниз позволила эффективно управлять затратами на операционную деятельность при снижении уровня спроса на некий стабильный уровень.

Качество эластичности весьма ценно не только для стартапов, но и для крупных компаний. Target, второй по величине ритейлер США, использует AWS для размещения своего сайта Target.com. 28 ноября 2008 г. (день «Чёрной пятницы») сайты других крупных ритейлеров демонстрировали крайне низкую

<sup>&</sup>lt;sup>5</sup> Иллюстрация http://www.eecs.berkeley.edu/Pubs/Tech Rpts/2009/EECS-2009-28.html

производительность или вообще были не способны функционировать в приемлемом режиме. Сайт target.com был на 50% медленнее, но обслуживал покупателей.

Другой пример: среди клиентов SalesForce.com встречаются компании с двумя пользователями, и компании с более чем 40,000 пользователями. Однако даже и не столь драматические ситуации колебаний спроса показывают одно из ключевых преимуществ Облачных Вычислений – риск ошибочного прогноза загрузки компьютерных ресурсов переносится от провайдера ПО-как-услуга на провайдера Облака. Провайдер Облака по разному может обрабатывать ситуацию риска. Например, при продаже почасовых порций ресурсов взимать большую удельную плату, нежели при продаже крупного пакета порций, например годовых контрактов.

## Расчёт экономической эффективности

Предположим, что провайдер Облачных Вычислений использует модель оплаты *pay-as-you-go* (пофакту-использования), когда потребитель платит пропорционально количеству времени и количеству использованных ресурсов.

Вторым предположением является то обстоятельство, что величина прибыли потребителя прямо пропорциональна общему числу потреблённых человеко-часов.

Существует несколько подходов к исчислению ценовых моделей для низкоуровневых, инфраструктурных сервисов. Анализ показывает, что модель исчисления по факту использования выглядит наиболее привлекательной из-за своей простоты и прозрачности для пользователей, как это и происходит в мире материальных полезностей, таких, как газ и электричество.

В своём простом виде, расчёт общей экономической оценки эффективности выглядит следующим образом:

ЧеловекоЧасы
$$_{\text{Облако}}$$
 × (Выручка – Затраты $_{\text{Облако}}$ ) ≥   
≥ ЧеловекоЧасы $_{\text{ДатаЦентр}}$  × (Выручка –  $\frac{3$ атраты $_{\text{Датацентр}}}{\text{Утилизация}}$ 

анализ показывает, что модель исчисления по факту использования выглядит наиболее привлекательной из-за своей простоты и прозрачности для пользователей

В левой части неравенства чистая прибыль умножается на число человеко-часов, что даёт оценку получаемой прибыли за указанное количество часов. В правой части производятся аналогичные вычисления для датацентра с фиксированной мощностью, с учётом значения средней утилизации использования оборудования в условиях пиковой и обычной загрузки. Большее значение левой или правой части неравенства соответствует возможности получения большей прибыли от использования Облачных вычислений или традиционного датацентра.

Очевидно, что если Утилизация = 1 (оборудование датацентра утилизируется на 100%), правая и левая часть неравенства выглядят идентичными. Однако теория очередей гласит, что при утилизации, стремящейся к единице, время ответа системы на запросы стремится к бесконечности. На практике, максимальная утилизация оборудования датацентра, без видимого снижения времени ответа на запросы пользовательским сервисом, составляет 0,6-0,8.

В условиях датацентра мы должны заложить этот резерв (0,4-0,2) на функционирование собственно самого датацентра. Учёт этого резерва — причина использования термина «pay-as-you-go», а не термина «аренда», в модели Облачных вычислений. Аренда включает резерв, в то время как «pay-as-you-go» — не включает. Аренда 100 мегабитного канала Интернет означает, что практическая пропускная способность составит порядка 60-80 мбит/сек. Аренда канала означает одну плату, плата за переданные мегабайты — другую.

Уравнение формализует и показывает один из основных общих элементов во всех наших примерах – возможность контролировать стоимость использования сервиса с детализацией до пользователя и до часа.

#### Выводы

В *Примере 1* стоимость пользователя-часа, без эластичности, достаточно высока из-за высокого уровня простоя ресурсов – и не меняется от уровня загрузки. Точно также при переоценке спроса на сервис, когда наше оборудование опять простаивает, стоимость пользователя-часа остаётся высокой.

В Примере 2, стоимость пользователя-часа увели-
чивается в результате недооценки спроса, что ведёт
низкому уровню обслуживания, следовательно, к
оттоку пользователей с сайта. В этом случае коли-
чество часов остаётся тем же, но количество пользо-
вателей уменьшается на количество ушедших, и не
вернувшихся пользователей.
Description of the second of t

Эти соображения в случае нетривиальных случаев колебания спроса, и соответственно, нагрузки иллюстрируют фундаментальные ограничения предыдущей до Облачной модели покупки сервисов.

#### Makarov S.V.

Deputy Director of the Centre "Innovative Educational Technologies" of The Graduate School of Corporate Management in The Academy of National Economy under the Government of the Russian Federation

# **Economics of cloud computing**

Abstract

In the article the author shows new approach to the deeper economic essence of the cloud computing model "pay-as-you-go", which may be interpret a principle of consumption of service – "payment only for use". It is also taken into account that in the cloud computing model user pays for the real use of utitity, but in the model of traditional data centre – for the use of resources of equipment.

**Keywords:** cloud computing model, economic efficiency of cloud computing, elasticity of cloud computing, traditional data centers