УДК 519.872

Разработка модели и алгоритмов оценки качества обслуживания системы облачных вычислений.

А.О. Волков

Московский физико-технический институт (государственный университет)
Институт радиотехники и электроники им. В.А. Котельникова РАН

В настоящее время системы облачных вычислений содержат десятки тысяч серверов, каждый из которых может включать в себя виртуальные машины с несколькими десятками ядер [1]. Кроме того, следует учитывать случайный характер распределения поступающих запросов и динамическое изменение рабочей нагрузки во времени для подобного рода систем. Для обеспечения приемлемого качества обслуживания (QoS, Quality of Service) в вышеперечисленных условиях, система облачных услуг должна быть готова к одновременному обслуживанию нескольких запросов. Все это требует точного и хорошо адаптированного механизма анализа производительности предоставляемого сервиса по модели SaaS (Software as a Service). Построенная математическая модель поступления и обслуживания заявок для системы облачных вычислений учитывает все вышеупомянутые особенности, модель представлена на рис. 1.

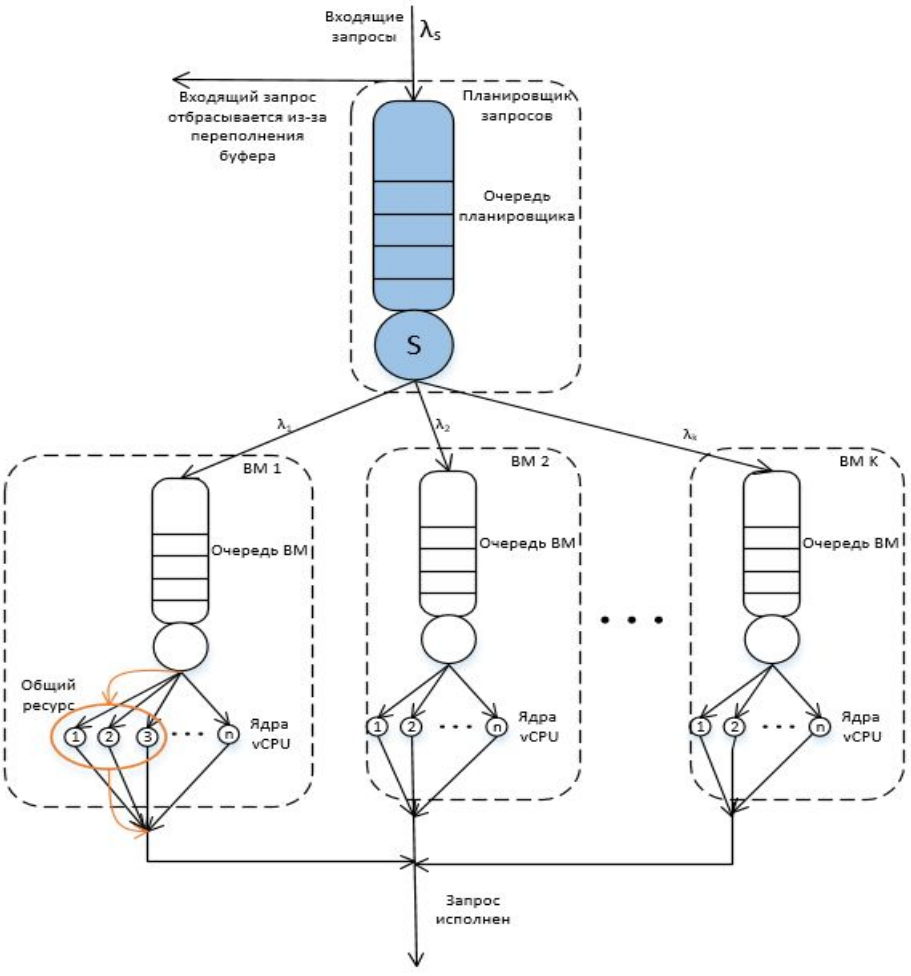

Рис. 1 Модель массового обслуживания системы облачных вычислений.

В облачных системах ресурсом обслуживания поступающих заявок является скорость обработки заявок в облаке [2]. В рамках модели SaaS клиенту предоставляется доступ к готовым приложениям и сервисам, которые работают в облаке провайдера и полностью обслуживаются

им. Клиент может заказать один или несколько облачных сервисов из k доступных. Облако обрабатывает заказанные сервисы с помощью определенного режима обслуживания [3]. В построенной модели для распределения запросов по ядрам процессора виртуальной машины (ВМ) используется режим PS (Processor Sharing).

Модель состоит из двух подсистем, первая из них включает в себя модель очереди планировщика, который балансирует нагрузку, а вторая – виртуальных машин. Поступление заявок в очередь планировщика осуществляется по закону Пуассона с интенсивностью λ_s. Очередь планировщика имеет конечную длину, то есть в случае её переполнения, входящий запрос отбрасывается. Планировщик распределяет запросы по локальным очередям, которые представлены экземплярами виртуальных машин, с вероятностью $\frac{1}{MN}$, где M – число виртуальных машин, развернутых в каждом сервере, а N – число серверов. В построенной модели предполагается, что каждая виртуальная машина имеет, по крайней мере, один виртуальный процессор (vCPU, virtual central processing unit) с одним или несколькими вычислительными ядрами, доступ к которым предоставляется по режиму Processor Sharing. В реальных облачных системах многоядерная архитектура позволяет распараллелить выполняемые запросы. Ввиду вышесказанного предполагается, что каждая виртуальная машина работает как сервер обслуживания, который представляет собой многоядерный vCPU, имеющий n идентичных ядер. Время выполнения запроса на ядрах ВМ является независимой случайной величиной со средним значением $\frac{1}{\mu}$, где μ – интенсивность обслуживания поступающих заявок.

При помощи построенной модели были получены формулы для ключевых показателей качества обслуживания и производительности системы облачных вычислений в модели SaaS с общим доступом, таких как: время отклика системы, время ожидания, загрузка системы и пропускная способность. Вышеперечисленные показатели QoS позволяют определить необходимое количество вычислительных ресурсов для развертывания приложений в системах облачных услуг, в том числе при пиковых значениях рабочей нагрузки. Разработанная модель и полученные параметры производительности могут быть также использованы для выявления «узких» мест в облаке. Кроме того, было показано, что системы, имеющие многоядерную архитектуру с разделением доступа к ресурсу, в условиях высокой нагрузки, показывают меньшее время ожидания, чем системы с одним или несколькими ядрами, в особенности это касается времени ожидания системы.

Литература

1. *B. G. Batista, J. C. Estrella, C. H. Ferreira, D. M. Filho и др.* Performance Evaluation of Resource Management in Cloud Computing Environments // PloS one. 2015.
2. *Степанов С.Н.* Теория телетрафика: концепции, модели, приложения / Серия "Теория и практика инфокоммуникаций" // Горячая линия – Телеком. 2015. С.868.
3. *D. Chitra Devi and V. Rhymend Uthariaraj.* Load balancing in cloud computing environment using Improved Weighted Round Robin Algorithm for nonpreemptive dependent tasks // Sci World J. 2016. С. 1–14