

ЗАДАЧА ПЕРВОНАЧАЛЬНОГО ВЫДЕЛЕНИЯ РЕСУРСОВ В ОБЛАЧНЫХ ВЫЧИСЛИТЕЛЬНЫХ СРЕДАХ НА ОСНОВЕ МЕТОДА АНАЛИЗА ИЕРАРХИЙ

УДК 004.051

Андрей Александрович Микрюков,
к.т.н., доцент, заведующий кафедрой
Автоматизированных систем обработ-
ки информации и управления ФГБОУ
ВПО «Московский государственный
университет экономики, статистики и
информатики»

Тел.: (495) 442-61-11

Эл. почта: AMikrakov@mesi.ru

Р.И. Хантимиров,

аспирант

ФГБОУ ВПО «Московский государс-
твенный университет экономики, ста-
тистики и информатики»

В статье рассмотрена задача повыше-
ния эффективности функционирования
облачной вычислительной среды на
этапе первоначального выделения
ресурсов с использованием метода
анализа иерархий (МАИ).

Ключевые слова: облачные вычисле-
ния, инфраструктура как сервис, IaaS,
выделение ресурсов, распределение
ресурсов, метод анализа иерархий.

Andrey A. Mikryukov,

Ph.D., associate professor, head of de-
partment of automated data processing
systems and management, Moscow State
University of Economics, Statistics and
Informatics

Tel.: (495) 442-61-11

E-mail: AMikrakov@mesi.ru

R. Hantimirov,

graduate student of Department of
automated information processing sys-
tems and management, Moscow State
University of Economics, Statistics and
Informatics

INITIAL RESOURCE PROVISIONING IN IAAS CLOUDS BASED ON THE ANALYTIC HIERARCHY PROCESS

The aim of this paper is to describe
one possible solution of a problem of
efficient initial resource provisioning in
the Infrastructure-as-a-Service cloud en-
vironments using the Analytic Hierarchy
Process.

Keywords: cloud computing, Infrastruc-
ture as a Service, IaaS, resource provi-
sioning, resource distribution, analytic
hierarchy process.

1. Введение

Сегодня при организации многопользовательских вычислительных процессов находит широкое применение концепция облачных вычислений, в основе которой лежит понятие облачной среды (далее – облака). Одной из популярных моделей предоставления услуг из облачной среды является модель «Инфраструктура как сервис» (IaaS – Infrastructure as a Service). Это обусловлено совокупностью следующих факторов:

1. Снижением капитальных затрат. При этом обеспечивается возможность снижения затрат на инфраструктуре вычислительной среды. Облачные вычисления позволяют добиться унификации используемых аппаратных средств и исключить единые точки отказа. Это делает возможным выбор в сторону недорогого и менее надежного оборудования, поскольку снижается критичность каждого отдельного элемента инфраструктуры для поддержания целевых показателей функционирования систем.

Благодаря унификации облака, потребитель зачастую вправе использовать бесплатные гипервизоры, решения для хранения данных и прочие компоненты, что, в свою очередь, дает возможность экономии на лицензиях.

Указанная модель позволяет снизить затраты на расширение облака. Модель IaaS построена так, что она может расширяться линейно, с использованием однотипных «строительных блоков». Более того, эти «строительные блоки» можно совершенствовать с течением времени без дополнительных затрат.

Снижением операционных затрат. При этом обеспечивается возможность снижения затрат на обслуживающий персонал, поскольку автоматизируются процессы предоставления, выделения, распределения ресурсов, биллинга, а также расширения и сжатия облака. Требуется значительно меньше трудозатрат на администрирование и техническую поддержку облака и его пользователей в сравнении с традиционными инфраструктурами и средами виртуализации.

Учитывая описанные выше преимущества, на начальном этапе формирования облачного вычислительного ресурса находит широкое применение модель «Инфраструктура как сервис» как для внутреннего использования, так и для предоставления услуг.

Таким образом, внутри облака могут размещаться самые разнообразные компоненты – начиная от малонагруженного веб-сайта и заканчивая высоконагруженными базами данных, ERP-системами и пр. Очевидно, что перечисленные компоненты предъявляют абсолютно разные требования к каждому типу вычислительных ресурсов, следовательно, размещать их на одних и тех же хостах, хранилищах и фрагментах сети нельзя – это приведет к нерациональному использованию ресурсов, снижению производительности и дополнительным затратам.

В [3] предложен подход, позволяющий решить задачу равномерного распределения нагрузки на все серверы и средства хранения данных разной производительности облачной вычислительной среды, обеспечивая при этом наилучшие условия для функционирования приложений и, в то же время, эффективное использование имеющихся вычислительных ресурсов. Один из этапов предложенного подхода связан с рациональным выделением ресурсов при запуске экземпляра (виртуального или физического сервера в облаке, предоставляемого как сервис). Решение указанной задачи рассмотрено в данной статье.

2. Анализ существующих подходов к первоначальному выделению ресурсов в облаке

Существующие в облачных средах подходы к первоначальному выделению ресурсов можно разделить на три типа: выделение ресурсов администратором, планирование ресурсов диспетчером облака и планирование ресурсов средой виртуализации [3].

Специфика распределения ресурсов в облаке «Инфраструктура как сервис» берет свое начало в основных идеях, которые заложены в этой модели облачного сервиса. Они хорошо описаны в [1]. В частности, это:

Отсутствие у администратора информации о реальных потребностях приложений, находящихся внутри экземпляров (чаще всего – виртуальных машин);

Отсутствие у администратора возможности ручного назначения ресурсов экземплярам.

Таким образом, облачные среды имеют особенности, которые создают специфику распределения ресурсов в облаке. Некоторые элементы этой проблематики являются общими с проблемами систем виртуализации, а некоторые – имеют особенности, характерные для облака. На процесс распределения ресурсов в облаке влияют:

Различные классы оборудования и сочетания ресурсов. Одной из основных идей при создании облачной среды является снижение затрат на его создание и расширение. Очевидно, что в составе облачной среды могут использоваться разные классы оборудования – серверы с разным числом процессоров, системы хранения данных с разной производительностью, сетевое оборудование с разной пропускной способностью и т.д.

Неравномерная загрузка каждого из ресурсов. Процесс функционирования облачной вычислительной среды предполагает минимальное вмешательство администратора для поддержания его работоспособности. Чтобы снизить время, затрачиваемое администратором на обслуживание облачной среды, и обеспечить его наиболее эффективную работу, при распре-

делении ресурсов стоит учитывать не только емкость, но и текущую фактическую нагрузку на каждый элемент инфраструктуры – это касается не только процессорного времени и оперативной памяти, но и многих других параметров, таких как отклик дисковых ресурсов, сетевые задержки и т.п.

Отсутствие сведений о реальных потребностях приложений в ресурсах. Администратор облачной среды не знает, какие приложения и с какими потребностями функционируют внутри экземпляра. Поэтому необходим механизм, который бы без участия человека определял, на каких ресурсах будет эффективнее функционировать то или иное приложение, не создавая проблем для других приложений.

Различные потребности приложений к ресурсам. В рамках двух экземпляров с похожими или вовсе одинаковыми параметрами (количество процессорных ядер, объем ОЗУ, диск) могут находиться два абсолютно различных по потребностям приложения. Например, в рамках экземпляра с 4-мя процессорными ядрами и 16-ю Гб ОЗУ может находиться как веб-приложение, которое будет использовать 3–5% процессорного времени и не давать большой нагрузки на диск, так и высоконагруженная база данных CRM-системы, которая будет иметь огромные потребности в дисковом ресурсе и использовать до 80% процессорного времени. Определение класса, к которому относится приложение, противоречит идее облачной среды и приводит к неэффективному использованию ресурсов, поскольку реальные потребности в них неизвестны.

Несоответствие запрошенных ресурсов потребляемым. Зачастую выделение ресурсов для информационных систем делается с большим запасом и не соответствует фактическим потребностям приложений, их составляющих. Поставщик облачного вычислительного ресурса преследует цель максимально эффективного его использования, и конфигурация облачной среды должна учитывать, что, к примеру, из запрошенных 256-ти Гб оперативной памяти по факту может

использоваться не больше 10-ти Гб. Если тарификация учитывает только используемые ресурсы, то и поставщик облачных ресурсов должен соотносить это со своими затратами.

Отсутствие возможности ручного назначения ресурсов. Как уже отмечалось выше, в облачной среде не существует механизмов ручной привязки пользователем приложений к конкретным серверам или системам хранения данных, так как это сводит на нет идеи, лежащие в основе облачных вычислений. Администратору неизвестно, какие приложения находятся внутри экземпляров, и он не имеет возможности вручную распределять ресурсы, еще и потому, что при большом числе экземпляров это фактически невозможно.

Таким образом, облачные среды обладают рядом особенностей, не характерных для других моделей предоставления ресурсов, что вызывает необходимость совершенствования механизмов распределения ресурсов для указанных сред.

В качестве критерия эффективности первоначального выделения ресурса предложен показатель эффективности функционирования облака $E_{обл}$, представляющий собой комплексный показатель, включающий ряд частных показателей и отражающий эффективность выполнения экземпляров в рамках каждого типа облачных ресурсов (процессорное время, ресурсы оперативной памяти, дисковые ресурсы, ресурсы сети и т.д.). Максимальное значение показателя эффективности соответствует максимальному числу экземпляров, функционирующих без потери производительности из-за перегрузки облачного ресурса.

В статье рассмотрена задача организации первоначального выделения ресурсов в облаке таким образом, чтобы обеспечить максимальное значение показателя эффективности функционирования ресурсов облака $E_{обл}$ при наличии ограничений к системным требованиям, т.е.:

$$E_{обл} = F(E_{ЦП}, E_{ОЗУ}, E_{д}, E_{с}, T, P_p) \rightarrow \max \quad (1)$$

где:

$E_{ЦП}$ – процент экземпляров с удовлетворенной потребностью в ресурсах центрального процессора;

$E_{ОЗУ}$ – процент экземпляров с удовлетворенной потребностью в ресурсах памяти;
 $E_{д}$ – процент экземпляров с удовлетворенной потребностью в дисковых ресурсах;
 $E_{с}$ – процент экземпляров с удовлетворенной потребностью в сетевых ресурсах;
 P_p – вероятность перераспределения ресурсов с хоста;
 T – время функционирования облака.

При этом необходимо учесть ограничения:

$$R_{исп} < R_{имеющ} \quad (2)$$

где:

$R_{исп}$ – общий объем использованных ресурсов облака;
 $R_{имеющ}$ – общий объем имеющихся ресурсов облака.

На этапе первоначального выделения ресурсов задача решается путем нахождения хоста, обеспечивающего наименьшую вероятность перераспределения ресурсов P_p^{min} с хоста, на котором выполняется первоначальный запуск экземпляра (в частности, виртуальной машины), что позволяет в процессе последующего функционирования облачной среды наиболее эффективно использовать имеющиеся ресурсы.

3. Решение задачи с помощью метода анализа иерархий

Анализ поставленной задачи показал, что она относится к классу многокритериальных задач принятия решений и может быть решена с использованием математического аппарата метода анализа иерархий. Как известно, метод анализа иерархий Саати успешно применяется в разных областях науки и техники для решения подобных задач [2,3]. Применение метода анализа иерархий включает ряд последовательных шагов с учетом особенностей решаемой задачи и специфики облачных сред:

1) Облачная вычислительная среда предварительно подразделяется администратором на зоны, состоящие из схожих по параметрам хостов и хранилищ данных;

2) С учетом заданных для экземпляра ограничений (на число процессоров, объем памяти, ресурсы

хранения и пр.) выбирается наиболее подходящая ему зона.

Выполняется оценка показателей качества каждого хоста и хранилища данных, включенных в выбранную зону.

3) Выполняется выбор наиболее подходящего хоста на основе анализа альтернатив.

4) Подробно каждый шаг описан ниже.

После предварительного разделения облачной вычислительной среды на зоны на первом шаге, на следующем шаге осуществляется выбор зоны, соответствующей требованиям для экземпляра. Использование зон позволяет учесть ограничения по характеристикам для запускаемого экземпляра, чтобы выбрать его месторасположение в той части облака, где ресурсов будет заведомо достаточно для обеспечения функционирования приложений, выполняемых на экземпляре.

На третьем шаге для каждого хоста из зоны, определенной на предыдущем этапе, выполняется оценка

значений показателей качества [4], которыми описывается функционирование хоста и его компонентов (рис. 1).

Указанные показатели являются составными, каждый из которых зависит от ряда частных показателей:

– производительность (P): число ЦП ($N_{ЦП}$), число ядер ЦП ($N_{ядер}$), тактовая частота ЦП ($V_{ЦП}$), скорость работы сети ($V_{сеть}$);

– надежность (R): среднее время между сбоями ($M_{сбоя}$), физическое состояние оборудования ($Q_{сост}$); качество сборки оборудования ($Q_{сборки}$); уровень компетенций персонала для обслуживания оборудования ($Q_{квал}$).

– нагруженность (S): свободная область ОЗУ ($R_{своб}$), загрузка ЦП ($L_{ЦП}$), нагрузка на дисковую подсистему ($L_{диск}$), отклик дисковой подсистемы ($T_{диск}$), загрузка сети ($L_{сеть}$), число выполняемых экземпляров ($N_{экз}$).

При этом совокупность частных показателей, влияющих на комплексный показатель качества, может быть сведена в 3 группы:

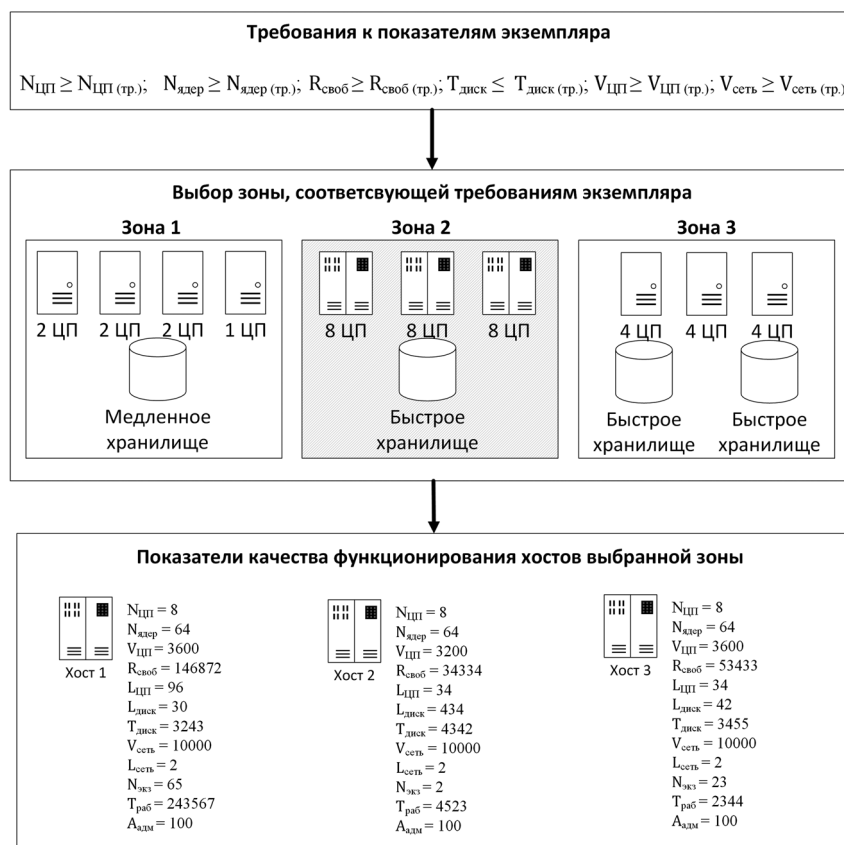


Рис. 1. Выбор зоны для запуска экземпляра, и оценка показателей качества функционирования хостов

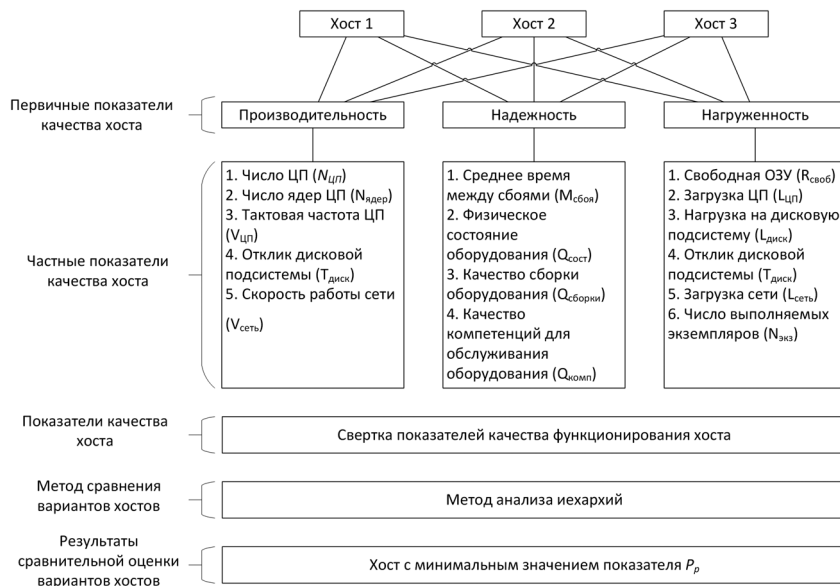


Рис. 2. Обобщенная схема алгоритма поиска хоста с минимальным показателем P_p

детерминированные (связанные с конфигурацией оборудования), стохастические (связанные с нагрузкой и временем функционирования) и качественные (часть указанных показателей, такие, как физическое состояние оборудования, качество его сборки и уровень компетенций обслуживающего персонала, не могут быть количественно определены и могут быть представлены в виде лингвистических переменных).

На четвертом шаге для выбора наиболее подходящего хоста для размещения экземпляра (запуска виртуальной машины на определенном хосте) выполняется анализ альтернатив на основе полученных значений показателей качества [5]. Каждый показатель имеет свой предварительно заданный вес, определенный экспертным путем для каждой зоны исходя из обобщенных потребностей клиентов и имеющегося оборудования.

В результате для размещения экземпляра выбирается хост, обеспечивающий минимальную вероятность последующего перераспределения ресурсов P_p^{min} . Структурная схема

решения указанной задачи представлена на рис. 2.

4. Вывод

Предложен подход, позволяющий обеспечить повышения эффективности функционирования облачной вычислительной среды, в которой реализована модель предоставления услуг «Инфраструктура как сервис», по заданному критерию эффективности. В качестве инструментального средства решения поставленной задачи использован метод анализа иерархий. Реализация предложенного подхода обеспечивает рациональное выделение ресурсов в облачной вычислительной среде при первоначальном запуске экземпляра (виртуального или физического сервера в облаке, предоставляемого как сервис).

Литература

1. The NIST Definition of Cloud Computing – <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
2. Саати Т. Принятие решений. Метод анализа иерархий. – М.: Радио и связь, 1993. – 320 с.

3. Хантимиров Р.И. Интеллектуальное планирование ресурсов в облачных средах на основе модели «Инфраструктура как сервис» // Сборник трудов молодежной научно-технической конференции «Информационно-коммуникационные технологии в экономике». – М.: МЭСИ, 2013. С. 378–385.

4. Анохин А. М. и др. Методы определения коэффициентов важности критериев // Автоматика и телемеханика. – М.: Москва, 1997. №8. – С. 3 – 35.

5. Микрюков А.А., Князев А.А. Об одном методе решения задачи многокритериального выбора // Сборник трудов II международной научно-практической конференции «Информационные технологии в образовании, науке и производстве», г. Серпухов, 2009. – С. 241–243.

References

1. The NIST Definition of Cloud Computing – <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>
2. Saati T. Prinyatie resheniy. Metod analiza ierarkhiy. – M.: Radio i svyaz', 1993. – 320 s.
3. Khantimirov R.I. Intellectual'noe planirovanie resursov v oblachnykh sredakh na osnove modeli «Infrastrukтура как сервис» // Sbornik trudov molodezhnoy nauchno-tekhnicheskoy konferentsii «Informatsionno-kommunikatsionnye tekhnologii v ekonomike». – M.: MESI, 2013. S. 378–385.
4. Anokhin A. M. i dr. Metody opredeleniya koeffitsientov vazhnosti kriteriev // Avtomatika i telemekhanika. – M.: Moskva, 1997. №8. – S. 3–35.
5. Mikryukov A.A., Knyazev A.A. Ob odnom metode resheniya zadachi mnogokriterial'nogo vybora // Sbornik trudov II mezhdunarodnoy nauchno-prakticheskoy konferentsii «Informatsionnye tekhnologii v obrazovanii, nauke i proizvodstve», g. Serpukhov, 2009. – S. 241–243.