Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра автоматизации систем вычислительных комплексов

Михеев Павел Алексеевич

Исследование эффективности использования физических ресурсов легковесными контейнерами в облачных системах.

Научный руководитель:
к.ф.-м.н.
В.А. Антоненко

Москва, 2018

Аннотация

Контейнер - изолированная группа процессов в операционной системе, имеющая доступ к ограниченному количеству ресурсов системы. Данная изоляция достигается за счет механизмов ядра ОС, таким образом, контейнеры используют это ядро, не тратя ресурсы на поддержку гостевого ядра.

Контейнеры обладают небольшим временем запуска, большой пропускной способностью, и показатель производительности вычислений в них ближе к показателю вычислений на физичеких ресурсах. Однако не существует систем, позволяющих исследовать производительность и деградацию производительности контейнеров при масштабировании и сравнить ее с производительностью и деградацией для виртуальных машин.

В данной работе предлагается использовать существующие методики измерения производительности виртуальных машин и применить их к легковесным контейнерам.

Целью работы является разработка системы измерения производительности и ее деградации у виртуальных машин и контейнеров, запущенных на физическом сервере.

Данная система позволит сравнить численно производительность контейнеров и виртуальных машин.

Введение

Облачные вычисления в современной информационной среде занимают все больше и больше пространства. Развитие открытого программного обеспечения по управлению облаками увеличивает количество тех, кто использует облачные системы как для использования внутри компании, так и для предоставления услуг клиентам. Одним из немногих сдерживающих факторов такого роста вовлеченности использования облачных вычислений является цена физического оборудования.

Традиционно облачные вычисления используют виртуальные машины как сущности, которые так или иначе предоставляются пользователю. Так, в случае облака, работающего по модели Infrastructure-as-a-Service (IaaS), пользователю предоставляется полный доступ к заказанной виртуальной машине, которую он настраивает под собственные требования. В облаках типа Platform-as-a-Service предоставляется доступ к интерфейсам некоторых приложений, запущенных в виртуальной машине. В облаках Software-as-a-Service пользователь запрашивает услугу, обработка которой происходит с помощью виртуальных машин, доступа к которым у пользователя нет.

Количество виртуальных машин, которое может быть запущено в облаке, напрямую зависит от количества физических ресурсов. Под количеством физических ресурсов в данной работе будет пониматься совокупное (по всем серверам) количество ядер процессоров, совокупное количество оперативной памяти и совокупный объем жестких дисков системы.

Пусть X - количество доступных физических ресурсов в облаке, а χ_i - количество ресурсов, требующихся для i-ой виртуальной машины. Тогда n - максимальное число виртуальных машин с такими требованиями, которое может быть запущено на данных ресурсах, если $\sum\limits_{i=1}^{n} \chi_i = X$.

Однако может возникать ситуация, когда количество запрашиваемых

виртуальных машин превосходит максимальное число n, определенное выше. При этом у владельца облака нет возможности или необходимости докупить оборудование, так как, к примеру, такая ситуация возникает редко. Важный момент, что ресурсы виртуальной машины могут быть использованы ее процессами не на 100% процентов. Тогда, если предоставить доступ к этим же физическим ресурсам или их части той виртуальной машине, номер которой превосходит n, то обе эти виртуальные машины смогут осуществлять свои функции. В таком случае $\sum\limits_{i=1}^{n} \chi_i > X$, но при этом все виртуальные машины размещены.

Введем коэффициент $\theta = \frac{\sum\limits_{i} \chi_i}{X}$ - коэффициент перекрытия (overlap), являющийся отношением суммы ресурсов, требуемых виртуальным машинам системы, к физическим ресурсам системы.

Разумеется, при увеличении коэффициента перекрытия производительность процессов в виртуальных машинах падает, так как при использовании одних физических ресурсов (в первую очередь, ядер процессора) разными виртуальными машинами исполнение машин одновременно будет невозможно. Из-за этого будет наблюдаться деградация производительности виртуальных машин при увеличении коэффициента перекрытия.

Еще одной быстро занявшей рынок технологией стала легковесная виртуализация. В этой технологии виртуализации гостевая операционная система отсутствует, а все процессы запускаются в рамках ядра хостовой операционной системы. Изоляция подобных процессов друг от друга и ограничение доступных им ресурсов достигается за счет специальных механизмов ядра. Такие процессы и их потомки с наложенными на них ограничениями называются контейнерами. Так же как и процессы в виртуальной машине изолированы от процессов другой виртуальной машины, процессы в контейнере изолированы от процессов других контейнеров. Отличие заключается в том, что системные вызовы в вирту-

альной машине идут в операционную систему машины, тогда как системные вызовы контейнеров идут напрямую к ядру операционной системы, в которой данный контейнер запущен.

Отсутствие необходимости поддерживать гостевую операционную систему обладает как достоинствами, так и недостатками. Подробнее о них будет сказано ниже.

Контейнеры могут быть использованы в облачных системах так же, как и виртуальные машины. Они потребляют некоторое количество ресурсов, ресурсы могут разделять между несколькими контейнерами, а при увеличении коэффициента перекрытия будет наблюдаться деградация производительности.

В данной работе предлагается исследовать и сравнить производительность виртуальных машин и контейнеров, а так же деградацию производительности машин и контейнеров при увеличении коэффициента перекрытия. Для этого была разработана система, способная запускать виртуальные машины или контейнеры и запускать в них приложения, позволяющие измерять производительность, а так же собирать результаты работы этих приложений.

Основной гипотезой данной работы является следующее: контейнеры обладают более высокой производительностью по сравнению с виртуальными машины, деградация производительности виртуальных машин при увеличении коэффициента перекрытия происходит быстрее, чем у контейнеров.

Глава 1.

Постановка задачи

Цель работы

Разработать и реализовать систему, позволяющую сравнить производительность виртуальных машин и контейнеров, а так же деградацию производительности виртуальных машин и контейнеров при увеличении коэффициента перекрытия.

С помощью разработанной системы провести эксперименты, позволяющие проверить следующую гипотезу: **контейнеры обладают более высокой производительностью по сравнению с виртуальными машины, деградация производительности виртуальных машин при увеличении коэффициента перекрытия происходит быстрее, чем у контейнеров.**

План решения задачи

1. Сравнить различные технологии виртуализации.

2. Составить обзор существующих решений, с помощью которого выделить методики оценки производительности виртуальных машин.

3. Выбрать из предыдущего обзора методику оценки производительно-

сти или разработать свою, которую возможно применить для оценки производительности контейнеров.

4. Разработать систему, реализующую методику оценки производительности из предыдущего пункта, позволяющий численно оценить производительность и ее деградацию при увеличении коэффициента перекрытия в случае виртуальных машин и контейнеров.

5. С помощью разработанной системы провести эксперименты, позволяющие проверить следующую гипотезу: **контейнеры обладают более высокой производительностью по сравнению с виртуальными машины, деградация производительности виртуальных машин при увеличении коэффициента перекрытия происходит быстрее, чем у контейнеров.**

Глава 2.

Технологии виртуализации

В данном разделе под хостовой операционной системой будет пониматься система, в которой могут быть запущены гостевые операционные системы. Гостевые операцинные системы - это те системы, которые видят лишь свое изолированное окружение, и которые не могут быть осведомлены о наличии других гостевых систем, кроме как через сеть. Под виртуальной сущностью будет пониматься тот процесс в хостовой операцинной системе, который исполняет вычисления гостевой операционной системы.

Гипервизор - специализированное программное обеспечение, которое занимается управлением гостевыми операционными системами: запуском, остановкой, наблюдением и выделением ресурсов. Гипервызиоры бывают двух типов:

1. Нативные гипервизоры, которые запускаются напрямую на оборудовании хоста, контролируют это оборудования и осуществляют наблюдение за гостевыми операционными системами.

2. Гипервизоры, которые запускаются поверх операционной системы хоста и осуществляют мониторинг гостевой операционной системы.

Виртуализация - такой подход к организации вычислений, при котором каждая виртуализированная сущность изолирована от других, при-

чем ей может быть доступна лишь часть общих ресурсов. В данном разделе будет расмматриваться виртуализация центрального процессора, то есть каким образом возможно исполнять гостевую операционную систему изолированно на центральном процессоре хостовой системы.

Существует четыре основных вида виртуализации: полная вирутализация, паравиртуализация, аппаратная виртуализация и легковесная виртуализация. Далее будут рассмотрены каждый из этих видов, а так же их достоинства, недостатки и применимость в облачных системах.

2.1. Полная виртуализация

Данный вид виртуализации является исторически первым. Основная его особенность заключается в том, что гостевая операционная системы полностью отделяется от управления инфраструктурой хоста. Гостевая ОС не требует никаких изменений, и не осведомлена, что запущена в виртуальном окружении.

Как следует из названия, данный вид виртуализации позволяет запустить любую гостевую операционную систему в любой хостовой. Единственное требование - это наличие динамического транслятора из машинного языка архитектуры, с которой работает гостевая операционная система, в машиный язык хостовой архитектуры. В основе работы данного вида виртуализации лежит принцип динамической трансляции [1].

Преимуществом данного вида виртуализации является гипотетическая возможность запускать любую гостевую операционну систему.

При этом основное преимущесво оборачивается и основным недостатком. При современном разнообразии вычислиельной техники невозможно иметь трансляторы с любого машинного языка в любой. Но даже при наличии транслятора возможно, что скорость исполнения гостевого кода будет гораздо медленнее, чем в случае исполнения на настоящей ар-

хитектуры. Дополнительно к этим недостаткам добавляются сложность реализации динамической трансляции, связанной, к примеру, с неразличимостью команд и данных.

При этом существует ряд гипервизоров, поддерживающих полную виртуализацию. Примерами могут быть VMware ESXi [?], Microsoft Virtual Server [1]. Стоит отметить, что в силу закрытости данного програмнного обеспечения, их исследование в данной работе проивзведено не будет.

2.2. Паравиртуализация

Проблемой динамической трансляции является исполнение привилегированных инструкций гостевой операцинной системы. При обработке данных инструкций работа гостевой операцинной системы ухудшается.

Паравиртуализация пытается исправить данный недостаток. Это подход подразумевает модификацию гостевой операционной системы таким образом, чтобы исполнение привилегированных инструкций было оформлено как системный вызов в хостовую операционную систему. При этом при использовании паравиртуализации производительность гостевой операцинной системы выше, чем при использовании полной виртуализации [3].

Основным же недостатком данного вида виртуализации является необходимость модификации ядра гостевой системы. Это уменьшает количество операцинных систем, которые могут быть виртуализированны с помощью данного подхода.

Наиболее популярные гипервизоры, поддерживающие паравиртуализацию: Xen [4] и VMware [1]. Однако в данной работе они не будут рассматриваться. Это связанно с тем, что технология Xen практичски не поддерживается сообществом, и, как следствие, не исользуется в облачных системах. Продукты же VMware являются закрытыми, что так же

не позволяет их использовать.

2.3. Аппаратная виртуализация

Подход аппаратной реализации подразумевает совпадение машинного языка гсотевой системы и хостовой. При этом архитектура процессора хостовой операцинной системы должна поддерживать аппартную виртуализацию, то есть исполнять код гостевой системы так, будто бы это код хостовой. Привилегированные вызовы автоматически отслеживаются гипервизором, и при обнаружении исполняются на процессоре напрямую [1].

Такой подход позволяет решить проблему необходимости модификации ядра для повышения производительности гостевой системы, а так же использование полной виртуализации, хоть и требует специфической архитектуры процессора хостовой системы. Однако данная технология есть почти во всех современных процессорах, что позволяет использовать данный вид виртуализации в облачных системах.

Аппаратная виртуализация поддерживается множеством гипервизоров. Основные представители - это VirtualBox [5], KVM [6], VMware ESXi [2]. В данной работе будет рассматривать гипервизор KVM, который является модулем ядра Linux.

2.4. Общие достоинства и недостатки

Такой подход имеет ряд достоинств и недостатков. С одной стороны, поддержка гостевого ядра расширяет набор возможных операцинных систем, которые можно запустить. С другой стороны, поддержка гостевого ядра требует ресурсов. Так, гостевая операционная система будет потреблять часть выделенной оперативной памяти, часть выделенного

Рис. 2.1. Схема аппаратной виртуализации

дискового пространства и часть ресурсов процессора. При запуске виртуальной машины будет тратиться время на запуск ядра ОС. Оперативная память, выделяемая гипервизором под виртуальную машину, чаще выделяется непрерывным участком, что может приводить к внутренней фрагментации памяти.

2.5. Виртуализация на уровне операционной системы

Другим подходом является виртуализация на уровне ОС.

Идея этого подхода такова: пусть операционная система способна выделять своим процессам (или группе процессов) и их потомкам некоторое подмножество своих ресурсов. Пусть так же ОС способна выделять для каждого такого подмножества ресурсов свои множества идентификаторов процессов. Тогда для каждого подмножества ресурсов возможно запускать процессы, которые в общем случае могут быть запущены только в единичном экземпляре (в первую очередь, это процесс init ядра Linux). Таким образом можно изолировать подгруппы физических ресурсов, имея возможность запустить на каждой подгруппе свой экземпляр операционной системы. Однако стоит отметить, что ядро этих ОС общее, то есть то, которая использовала хостовая ОС.

Другим названием для этого подхода служит "контейнеризация", "легковесная виртуализация", или "контейнерная виртуализация" а каждую

подгруппу ресурсов со своим множеством процессов называют "контейнеры".

Изначально этот подход был реализован в ядре Linux. Модуль control groups (cgroups) ядра позволял выделять подгруппы ресурсов, а модуль namespaces - выделять свои иерархии идентификаторов процессов. На базе этого подхода реализовано несколько видов контейнеров - в первую очередь Linux Containers(далее LXC).

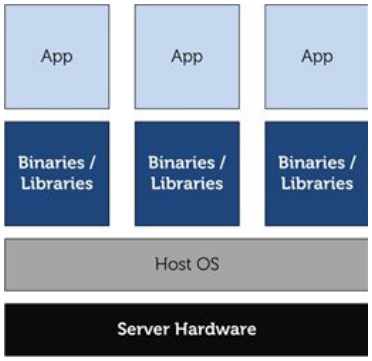

Рис. 2.2. Схема виртуализации на уровне ОС
labelконтейнеры

Выделим необходимые множества ресурсов ОС, требующих выделения подгрупп:

- Точка монтирования: разные контейнеры должны монтироваться в разные точки файловой системы хостовой ОС. Иначе действия контейнеров над файлами не будут синхронизированы.

- Network namespace: это необходимо для обеспечения полной изоляции на уровне сокетов, IP адресов и портов от соседа. Таким образом каждый контейнер имеет строго зафиксированный отдельный IP адрес, а также пространство сокетов, портов и роутинг-таблицу.

- IPC namespace: изолирует семафоры, очереди, мьютексы, shm память для IPC и IPC Sys V в таком виде, что каждый контейнер видит только свои IPC ресурсы и ничьи более.

- PID namespace: изолирует идентификаторы процессов в разных контейнерах друг от друга и от ОС, в которой запущен этот контейнер.
- UTS: позволяет выдавать уникальные hostname для каждого контейнера так, чтобы они не совпадали друг с другом и именем ОС, в которой контейнеры запущены.

Так же между контейнерами надо разделять следующие физические ресурсы: процессор, память и жесткий диск. Для реализации разграничения использования ресурсов используются следующие cgroups: cpu, memory и blkio. Например, с помощью cpu можно ограничить число ядер, которые используются в этой cgroup. Стоит отметить, что при контейнерной виртуализации для каждого контейнера используется отдельный набор пространств имен (namespace) и отдельный набор cgroups, то есть разделение их между контейнерами не используется.

Типичными представителями контейнеров легковесной виртуализации в ОС Linux является Linux Containers [7] и OpenVZ [8]. Однако, в отличии от LXC, OpenVZ использует свой набор утилит для контейнеризации, пусть общая идея такая же, как описано выше. Чтобы работать с OpenVZ, необходима специальная настройка ядра под эти утилиты [8]. В то время как LXC работает в рамках стандартного ядра Linux.

Важной особенностью является тот факт, что все эти хосты работают внутри одной ОС, то есть все контейнеры, запущенные в рамках одной ОС, используют ядро ОС как свое ядро. Таким образом, нет возможности изменить ядро ОС для одного контейнера, не затронув остальные.

2.6. Контейнеризация приложений

Рассмотрим ситуацию, в которой пользователь хочет запустить новое приложение на своей машине. Однако скачивание, установка и настройка системы для работы с этим приложением занимает много времени и

усилий, а некоторые дополнительные пакеты имеют несовместимости с существующими на машине пользователя. Возникает необходимость создания изолированного окружения для приложения, которое содержит все необходимые зависимости и настройки, при этом это окружение недоступно приложениям извне.

Контейнеризация приложения - это упаковка приложения и всех его зависимостей в легковесный контейнер и последующая настройка этого контейнера. Примером может послужить перенос ftp-сервера с машины на машину. После установки в LXC-контейнер самого сервера, необходимо настроить проброс 21 порта контейнера на 21 порт машины пользователя. Таким образом, когда пользователь скачает такой LXC-контейнер и запустит его, то сразу получит работающий FTP-сервер, слушающий 21 порт.

Следующим шагом служит автоматизация работы с LXC-контейнером, то есть необходимые действия по упаковке и запуску осуществляет не пользователь, а программа.

Docker [9] - средство для контейнеризации приложения со всеми его зависимостями, обеспечивающее API высокого уровня к LXC-контейнерам. Приложение упаковывается в Docker-контейнер, файловая система которого может переноситься с машины на машину.

Docker позволяет модифицировать Docker-контейнеры, добавляя в них новые приложения. При этом создается новый контейнер.

Последнее свойство нуждается в пояснении своей реализации. Если в LXC-контейнере используется обычная файловая система, то файловая система Docker-контейнера образована "слоями" из файловых систем - то есть в Docker-контейнере добавление файла в контейнер осуществляется путем записи нового слоя, который содержит файл, поверх старых слоев. В момент запуска Docker-контейнера старые слои интерпретируются как единая файловая система с помощью средства union-filesystem

(разработка команды создателей Docker), причем эта единая система доступна только для чтения. Все изменения записываются поверх нее в новый слой (стоит отметить, что в документации Docker нет пояснений по поводу реализации этой системы). Такой подход позволяет использовать чужие Docker-контейнеры как основу для новых.

Таким образом, Docker - удобная система для управления приложениями. Так же как и LXC-контейнеры, Docker-контейнеры, запущенные в рамках одной ОС, обладают общим ядром. Это приводит к тем же ограничением на запуск приложения в контейнере, что и для LXC-контейнеров.

За управление Docker-контейнерами отвечает Docker-daemon, процесс, работающий в хостовой ОС. Он запускает необходимые контейнеры (то есть либо собирает новый LXC-контейнер, либо запускает копию уже существующего). Docker-daemon выдает IP-адрес контейнеру из диапазона, задаваемого IP-адресом и маской сети сетевого интерфейса, который указан в настройках Docker.

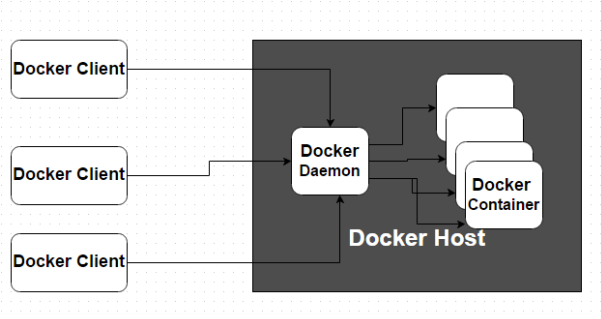

Рис. 2.3. Docker-демон

Сборка контейнера может производиться автоматически по Dockerfile - специальному файлу, в котором описана последовательность действий, необходимая для формирования нужного контейнера. Пользователь может сам написать подобный файл, а затем тот, кто получит такой файл, может сам собрать по нему нужный контейнер. При запуске Docker-контейнера возможно указать ограничения на размер доступной этому

контейнеру оперативной памяти и число ядер процессора, на которых можно исполнять приложения этого контейнера.

Стоит отметить, что Docker не единственное средство для контейнеризации приложения. Существует так же Rocket для Linux [10] и Drawbridge для Windows [11]. Однако наиболее развитым и активно поддерживаемым со стороны общественности является Docker.

Начиная с 2016 года, был разработан runC - описание поведения системы, производящей управление контейнерами. Хоть официально runC не является стандартом, но позволяет унифицировать описание работы контейнера, абстрагируясь от конкретных сущностей, и этому описанию стараются следовать все разработчики контейнеров. Однако родоначальником runC является Docker, то есть Docker работает по классической схеме runC [12]. Таким образом де-факто Docker становится стандартом контейнерной виртуализации.

Начиная с 2016 года, разработчики Windows и Docker анонсировали внедрение Docker в ядро Windows. Достигнуто это будет за счет того, что была добавлена возможность запуска Ubuntu в Windows за счет транслирования системных вызовов ядра Linux в системные вызовы ядра Windows. Аналогичные разработки ведутся и под MacOS.

2.7. Выводы

Наиболее часто используемыми технологиями в сфере облачных вычислений являются легковесная и аппаратная виртуализации. Первая технология позволяет управлять контейнерами, а вторая - виртуальными машинами.

В рамках данной работы будут рассматриваться гипервизор KVM для управления виртуальными машинами и система Docker для управления контейнерами приложений.

В KVM существует понятие виртуального процессора (vCPU). Эта абстракция описывает ядро, доступное виртуальной машине. Обычно виртуальной машине выделяются не все количество ядер хостовой системы,а лишь часть. При это по умолчанию в разные моменты времени виртуальная машина может как виртуальные ядра использовать разные ядра хостовой системы. Это контролирует гипервизор. В KVM присутсвует возможность закрепить конкретные ядра за конкретной машиной.

В Docker система выделения ресурсов процессора контейнерам ближе к системы выделения ресурсов для процессов. При это отсутсвует возможность ограничить число доступных контейнеру ядер, возможно лишь четко заификсировать, какими ядрами пользоваться контейнеру.

В KVM выделение оперативной памяти вирутальной машине происходит непрерывным блоком. То есть, даже если машина пользуется не всей доступной памятью, нет возможности передать неиспользуемую память другой машине.

Контейнеры обладают более гибкой системой управления оперативной памятью. Возможно задать лишь верхнюю границу выделяемой контейнеру памяти. При этом неиспользованная память может быть передана другому контейнеру.

Файловая система виртуальной машины является непрерывным файлом в хостовой операционной системе. Контейнеры же монтируют свою файловую систему в файловую систему хоста.

LXC-контейнеры потребляют меньше ресурсов, чем виртуальные машины, так как нет необходимости поддерживать ядро гостевой ОС, и при этом позволяют запускать изолированные приложения. Запуск контейнеров происходит быстрее, чем запуск виртуальной машины. [13].

Глава 3.

Обзор существующих решений

Привет, git. Почему ты не работаешь?

Литература

[1] Hyungro Lee. Virtualization Basics: Understanding Techniques and Fundamentals. // School of Informatics and Computing, Indiana University, 2014.

[2] Homepage of VMWare ESXi. http://www.vmware.com/ru/products/esxi-and-esx.html (дата обращения 01.09.2017)

[3] Hasan Fayyad-Kazan, Luc Perneel, Martin Timmerman. Full and Para-Virtualization with Xen: A Performance Comparison. // Journal of Emerging Trends in Computing and Information Sciences, 2013.

[4] Homepage of Xen. https://www.xenproject.org (дата обращения 01.09.2017)

[5] Homepage of VirtualBox. https://www.virtualbox.org (дата обращения 01.09.2017)

[6] Homepage of KVM. https://www.linux-kvm.org (дата обращения 01.09.2017)

[7] Homepage of LXC. URL: https://linuxcontainers.org/ru/ (дата обращения 01.09.2017)

[8] Homepage of OpenVZ. URL: https://openvz.org/ (дата обращения 31.10.2017)

[9] Homepage of Docker. URL: https://docs.docker.com/ (дата обращения 01.09.2017)

[10] Homepage of Rocket. URL: https://coreos.com/rkt/docs/latest/ (дата обращения 31.10.2017)

[11] Homepage of DrawBridge. URL: http://research.microsoft.com/en-us/projects/drawbridge/ (дата обращения 31.10.2017)

[12] Homepage of runC. URL: https://runc.io/ (дата обращения 30.04.2017)

[13] Михеев Павел. Разработка и реализация системы масштабирования виртуальных сетевых функции с помощью легковесной виртуализации. // МГУ, 2017.

[14] George Kousiouris, Tommaso Cucinotta, Theodora Varvarigoua. The effects of scheduling, workload type and consolidation scenarios on virtual machine performance and their prediction through optimized artificial neural networks. // The Journal of Systems and Software, 2011.

[15] Anton Beloglazov, Rajkumar Buyya. Optimal online deterministic algorithms and adaptive heuristics for energy and performance efficient dynamic consolidation of virtual machines in Cloud data centers. // Wiley Online Library, 2011.

[16] Haikun Liu, Hai Jin, Cheng-Zhong Xu, Xiaofei Liao. Performance and energy modeling for live migration of virtual machines. // Springer US, 2013.