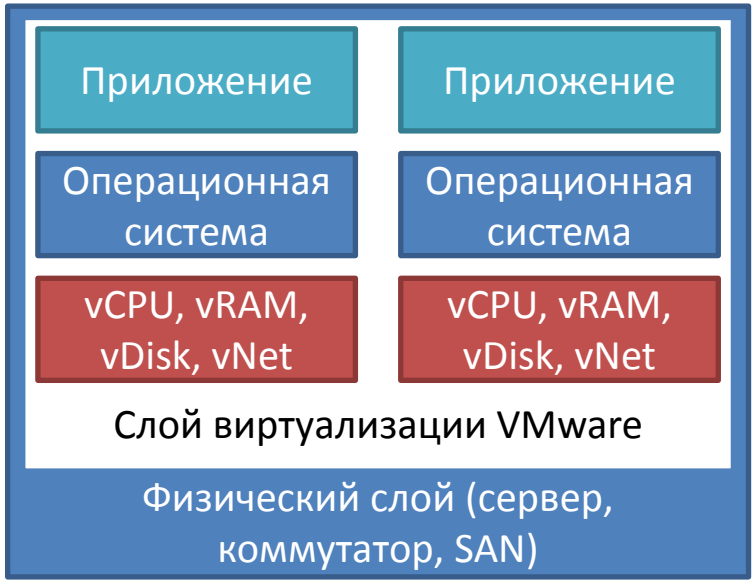
Настройка и профилирование виртуальной инфраструктуры VMware для интенсивного ввода/вывода PostgreSQL

Докладчик: Смолин Александр Сергеевич

PostrgreSQL в виртуальной инфраструктуре VMware

- Для 80% случаев достаточно применения стандартных настроек виртуализации VMware
- Для остальных 20% повышение производительности достигается за счет снижения накладных расходов:
  - эксклюзивное резервирование физических ресурсов
  - обход уровней виртуализации для устранения накладных расходов на дополнительную обработку
  - настройка слоев виртуализации

Общая модель виртуализации

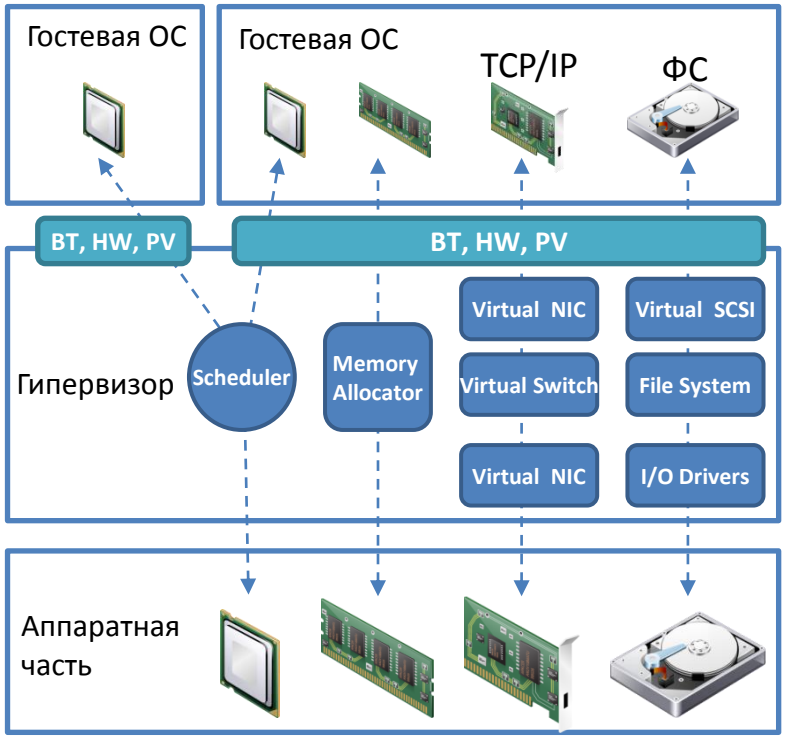


Слой приложений
- Библиотеки и сервисы

Слой гостевой операционной системы
- Локальный планировщик процессов
- Менеджер памяти
- Локальная файловая система

Слой виртуализации
- Планировщик процессора
- Менеджер ресурсов
- Драйвера устройств
- Стек ввода-вывода
- Файловая система
- QoS сети
- Межсетевой экран
- Управление электропитанием
- Управление неисправностями
- Наблюдение за производительностью

Архитектура VMware ESXi

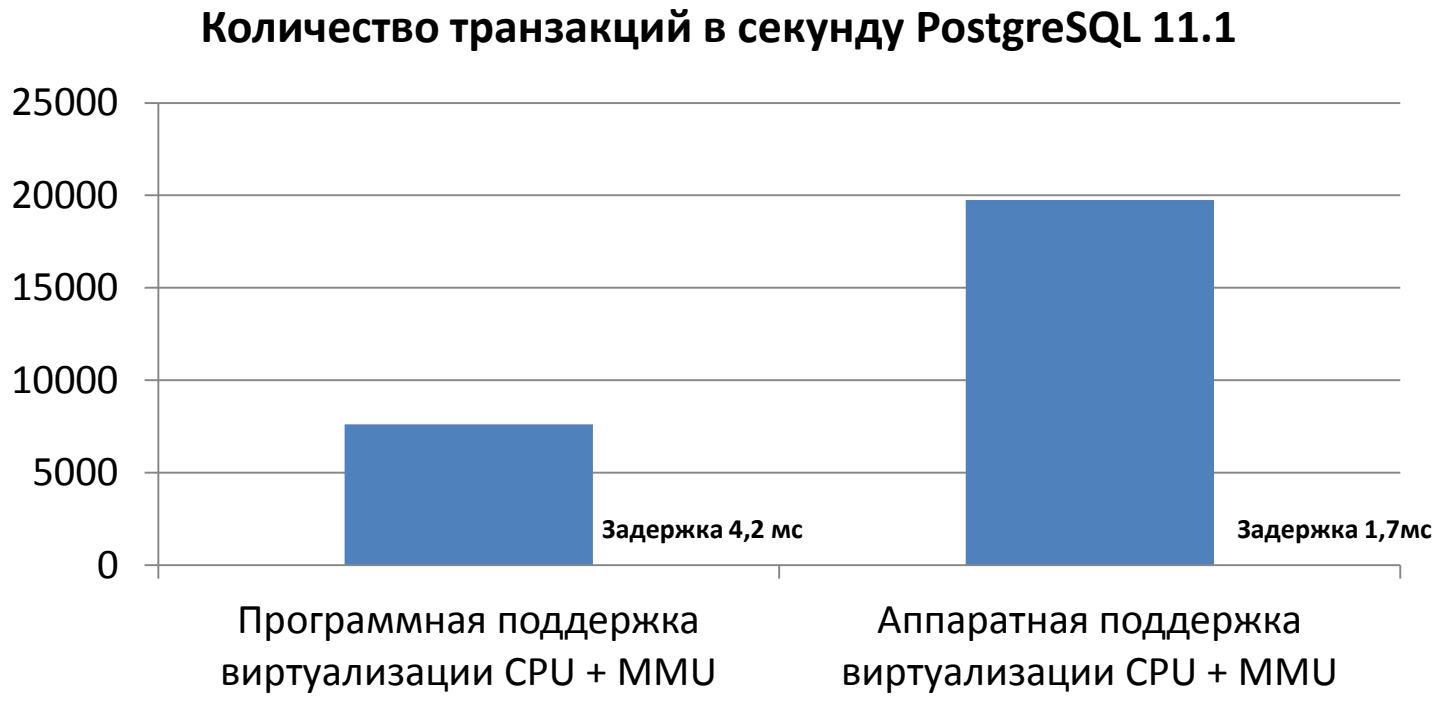


- Процессором управляет планировщик гипервизора
- Гипервизор поддерживает технологии:
  - Двоичная трансляция (BT)
  - Аппаратная виртуализация (HW)
  - Паравиртуализация (PV)
- Оперативная память распределяется гипервизором
- Сетевые устройства и устройства ввода-вывода эмулируются и проксируются через собственные драйверы устройств

Способы увеличения производительности

- Процессор
  - Аппаратная поддержка виртуализации (Intel VT/AMD-V)
- Оперативная память
  - Аппаратная поддержка виртуализации оперативной памяти
- Устройства ввода-вывода
  - Паравиртуализированные устройства
  - Проброс в виртуальную машину PCI Express устройств (VT-d/IOMMU)
    - Не будет работать VMotion и Storage VMotion, FT, snapshot и приостановка работы виртуальной машины

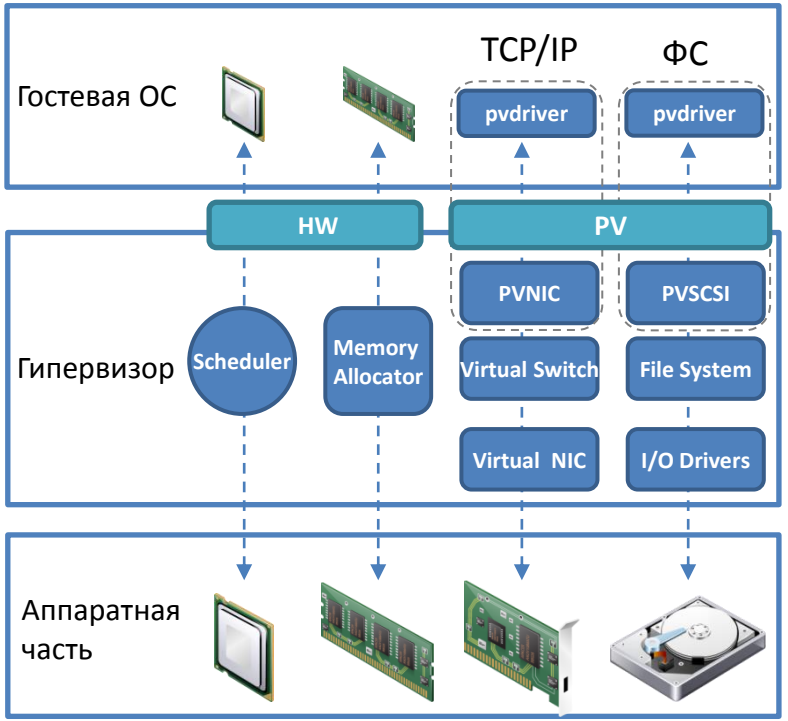
Сравнение аппаратной и программной поддержки виртуализации CPU и MMU



HW: Intel Xeon Gold 6152 - 2 шт, 190 GB RAM ; BM: 8 CPU, 32 Gb RAM, RAID 10; pgbench -s 100 -c 32 -j 8 -N -T 300 test

Аппаратная поддержка виртуализации и паравиртуализация

- Виртуальный дисковый контроллер Paravirtual SCSI (PVSCSI) и сетевой адаптер VMXNET3 позволяют на 30% снизить накладные расходы на CPU



Причины проблем производительности

- Слишком большое количество vCPU снижает эффективность работы и увеличивает накладные расходы
- Борьба за дисковый ввод/вывод при записи WAL, логов или во время других интенсивных дисковых операций
- Перерасход оперативной памяти
- Сетевые задержки
- Созданные снимки состояния системы (snapshot)

Проверка производительности CPU

- Используйте утилиту esxtop
- проверьте значение **load average**
- значение 1.0 означает полную утилизацию всех вычислительных ядер CPU, значение 2.0 и более означает перегруженность
- поле %READY отображает процент времени в течении которого виртуальная машина ожидала запуска на CPU

- Если загрузка ЦП не вызвана ограничениями установленными для CPU, увеличьте количество физических процессоров или уменьшите количество виртуальных CPU выделенных хосту ESXi, или уменьшите количество виртуальных машин на хосте

Дисковая система

- RAID 10 более эффективен для записи, чем RAID 5-6, меньше деградирует при отказе диска, но требует больше дисков
- Выполнить выравнивание диска. Веб-клиент vSphere начиная с версии ESXi 5.0 автоматически выравнивает разделы VMFS3, VMFS5 или VMFS6 по границе 1 МБ
- Не выровненный диск и не выровненная гостевая файловая система замедляют производительность на 30%
- Используйте паравиртуальный адаптер Paravirtual SCSI (PVSCSI)
- Тип диска Thick Provisioning Eager Zeroed обеспечивает наилучшую производительность и позволяет избежать накладных расходов при выделении дискового пространства
- Storage IO Control
- FC HBA предоставляет максимальную производительность и минимальные задержки
- Снимки состояния виртуальной машины снижают производительность
- Во время консолидации снимков состояния требуется много IOPS (надо учитывать)

Данные и WAL

- Данные и WAL должны находиться на разных физических дисках или LUN
- Предоставление виртуальной машине нескольких LUN подключенных к разным паравиртуальным SCSI адаптерам позволяет хосту ESXi предоставить больше запросов ввода/вывода к СХД. Это поможет повысить производительность, обеспечивая большую утилизацию СХД
- Если нет возможности разместить на разных дисках, сделайте разные виртуальные диски на одном физическом хранилище

Настройка адаптера PVSCSI и Host bus adapter (HBA)

- Нагрузки высокоинтенсивного ввода/вывода требуют большей глубины очереди адаптера PVSCSI

- Пример настройки HBA Emulex LPe16000 и CentOS 7:

  ESXi:

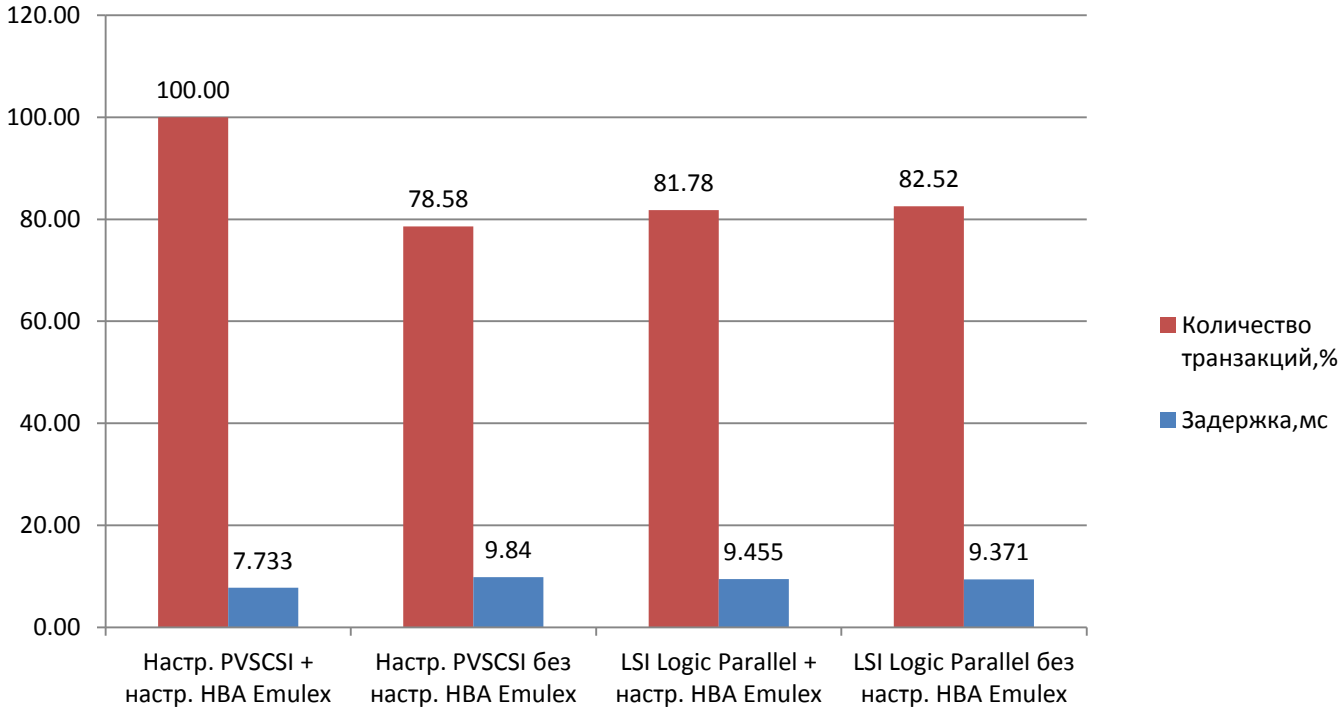
  ```
  esxcfg-module -s "lpfc_lun_queue_depth=128" lpfc
  ```

  CentOS 7:

  /etc/default/grub

  ```
  vmw_pvscsi.cmd_per_lun=254
  vmw_pvscsi.ring_pages=32
  ```

Сравнение настроек PVSCSI, LSI Logic Parallel и HBA Emulex



HW: Intel Xeon Gold 6152 - 2 шт, 190 GB RAM , HBA Emulex LightPulse LPe16000; BM: 16 CPU, 32 Gb RAM, RAID 10; pgbench -s 700 -c 200 -j 8 -N -T 300 test

Конфликт драйвера vmsync (только в Linux)

- Не используйте снимки состояния при высокой нагрузке ввода/вывода

- Во время выполнения снимков состояния при высокой нагрузке ввода/вывода PostgreSQL гостевая операционная система может перестать отвечать на запросы и внешне казаться зависшей
- Переход кластера corosync/pacemaker на резерв происходит во время создания или удаления/консолидации snapshot, а после окончания заморозки, отстрел старого мастера с помощью zstonith агента

Проверка производительности дисковой системы

- Используйте esxtop
- DAVG, средняя задержка (мс) на устройстве LUN
  - для WAL (<2-5 мс)
  - data (<10 мс)

- Проверьте максимальный ввод/вывод iometer, результаты можно сравнить с физической машиной подключенной к такой же системе хранения

- При использовании iSCSI и jumbo фреймов проверьте настройку размера фреймов на всем пути следования

Уменьшение задержки дисковой системы

- vSphere Flash Read Cache (vFRC)
  - Задержка минимизируется за счет использования SSD в качестве кэша чтения

Оперативная память

- Swap in rate и Swap out rate активность использования файла подкачки, должно быть 0
- Balloon должно быть 0
- N%L > 80 оптимально для NUMA

Операционная система

- Установите VMware Tools (драйвера и утилиты)
- Планируйте резервное копирование и вирусное сканирование в часы наименьшей загрузки
- Настроить синхронизацию времени гостевой операционной системы и хоста ESXi с NTP службой
- Отключите SELinux, если политика информационной безопасности позволяет

Отключите не используемые устройства

- Виртуальные устройства, такие как Floppy дисководы, CD-ROM, COM-порты и т.д. потребляют ресурсы
- Экранные заставки

Сеть

- Используйте серверного класса сетевые карты
- Можно включить Jumbo frame
- VMXNET3 для наилучшей производительности и наименьших накладных расходов
- Требует установленные VMware Tools
- Network IO Control

Проверка производительности сети

- Исключите проблему производительности CPU, может быть следствием
- Проверьте настройки ограничений для сети (Network I/O Control, traffic shaping)
- Счетчики droppedRx и droppedTx в графическом интерфейсе, или %DRPTX и %DRPRX в esxtop

Аппаратная часть

- Аппаратная часть должна быть совместима с VMware (http://vmware.com/go/hcl)
- Используйте, только серверные комплектующие
- Firmware актуальной версии
- Включите в BIOS:
  - Turbo Mode
  - Аппаратную поддержку виртуализации
  - Максимальную производительность в настройках энергосбережения

Проверка аппаратной поддержки виртуализации

esxcfg-info | grep HV

0 – недоступна

1 – доступна, но не поддерживается

2 – доступна, выключена в BIOS

3 – включена в BIOS и доступна

ПО для просмотра метрик

- esxtop – показывает информацию в реальном времени, похож на top
- resxtop – позволяет удаленно подключиться к хосту esxi или vcenter
- esxplot – графическое представление метрик
- lpar2rrd – импортирует из vCenter и хранит не усредненную информацию

Рекомендуемые ресурсы

VMware VROOM! Blog from VMware's performance team

http://blogs.vmware.com/performance

VMware Technical Papers

http://www.vmware.com/vmtn/resources

Large-scale workloads with intensive I/O

https://kb.vmware.com/s/article/2053145

A virtual machine can freeze under load when you take quiesced snapshots

https://kb.vmware.com/s/article/5962168

VMware Performance for Gurus

https://www.slideshare.net/rjmcdougall/usenix10-vmwareperf23

Спасибо за внимание!

Докладчик: Смолин Александр Сергеевич
smolinconf@gmail.com