Оценка потребления физической памяти в виртуальных машинах

Мелехова Анна Леонидовна, МФТИ, Parallels

Научный руководитель: доктор физ.-мат. наук, профессор, зав. Кафедрой теоретической и прикладной информатики МФТИ, проректор по науке университета Иннополис
Тормасов Александр Геннадьевич

Структура выступления

1. Постановка задачи оценки потребления физической памяти в виртуальных машинах
2. Актуальность задачи
3. Проведение оценки
    1. Методика получения данных
    2. Поиск корреляций среди событий виртуализации
    3. Проверка однородности данных от разных ОС
    4. Корректировка оценки с учетом обратной связи
4. Верификация полученной оценки

Постановка задачи (1)

Постановка задачи (1)

1. Виртуальная машина предоставляет виртуальное окружение времени исполнения
2. Ресурсы, предоставляемые в окружении, отображаются на ресурсы реальной физической машины
3. Заявленных виртуальных ресурсов в большинстве случаев меньше имеющихся физических ресурсов
4. Для корректного распределения физических ресурсов между виртуальными машинами требуется оценить потребность гостевой системы в ресурсе

Постановка задачи (2)

5. Потребление всех ресурсов кроме физической памяти прозрачно для гипервизора
    1. Гостевая ОС уведомляет о простое ЦПУ
    2. Устройства ввода/вывода не активны вне запроса
    3. Неиспользованная физическая память кэшируется ОС

Актуальность задачи (1)

1. Облачные инфраструктуры полагаются на автоматическое распределение ресурсов
2. Ресурсы находятся в режиме оверкоммита (переназначения)
3. Огромный аппарат для управления физической памятью виртуальных машин, но…

Актуальность задачи (2)

Подходы к управлению физической памятью виртуальной машины

1. Технология «воздушный шар»
2. Компрессия неиспользованной памяти
3. Слияние страниц (same page merging)
4. Применение алгоритмов по управлению виртуальной памяти процесса
    1. Гистограмма LRU
    2. Произвольный
    3. Старение (aging)
    4. WSClock

Актуальность задачи(2)

Технология «воздушный шар» (balloon)

Гипервизор сообщает размер воздушного шара своему агенту в гостевой ОС

Агент аллоцирует память и сообщает гипервизору адреса полученных физических страниц

Гипервизор отдает неиспользованные страницы другому потребителю памяти.

Актуальность задачи (3)

Недостатки существующих технологий

1. «Воздушный шар» -> снижение производительности + гостевые крэши
2. Компрессия -> размер выигрыша не стабилен
3. Слияние страниц -> размер выигрыша не стабилен
4. Алгоритмы виртуальной памяти процесса -> семантический зазор + неэффективный результат

И ни одна из этих технологий не способна предсказать потребление памяти даже на шаг вперед

Актуальность задачи (4)

Оценка потребления физической памяти - это

1. Правильный размер «воздушного шара»
2. Справедливое распределение ресурсов между ВМ на одной физической машине
3. Более эффективный менеджмент ресурсов в облачной инфраструктуре (между многими физическими машинами)

Актуальность задачи (5)

Почему не подходят классические методы «рабочего набора»

1. Семантический зазор
2. Игнорирование пласта знаний от гостевой ОС

Почему информации от гостевой ОС недостаточно

1. Free в Linux очень быстро достигает нуля
2. Свободные != не будет доступа

Моделирование

Почему эта идея родилась

- Сбор статистики по виртуализационным событиям не вносит накладных расходов
- События соотносятся с гостевыми паттернами
- Потребление памяти коррелирует с гостевыми паттернами
- Виртуализационные события коррелируют с потреблением памяти

Сбор статистики (1)

Сбор статистики(2)

Наблюдаемые виртуализационные события

- Количество страничных промахов ВМ
- Количество вытесненных страниц
- Количество различных страничных преобразований
- Частота переключения страничных преобразований
- Количество гостевых страничных промахов
- Частота доступа к регистру приоритета задачи
- Количество гостевых инструкций простоя
- Количество гостевых прерываний
- Процент времени, проведенный в госте
- Количество гостевых событий ввода/вывода
- Общее число виртуализационных событий

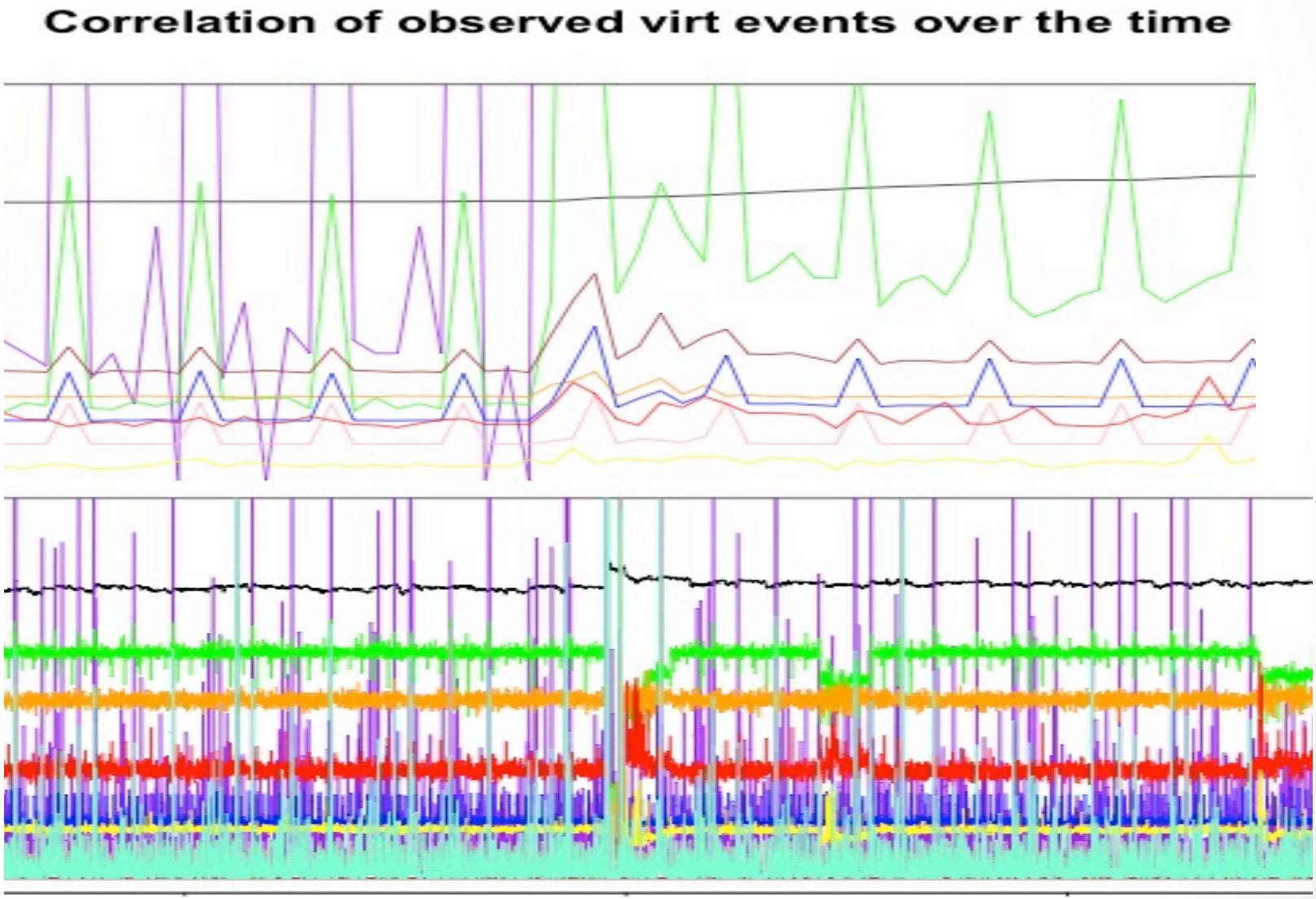
Сбор статистики(3)

Рабочая нагрузка для сбора выборки

- Простой/занятый цикл(busy loop)
- Тесты с дисковой активностью
- Тесты на переключение процессов/потоков
- Тесты с сетевой активностью
- Тесты-потребители памяти
- Инсталляция MS Office
- Симуляция работы в Office приложениях
- Тесты на 3D графику

Время сбора статистики – 28 часов. Одна выборка собирается 5 секунд.

Анализ корреляций(1)



Анализ корреляций(2)

- Размер рабочего набора растет при страничных промахах
- Рост рабочего набора идет постепенно даже в случае интенсивного использования памяти (w=f(t)), т.е. мы можем предсказать дельту, но не абсолютный показатель.
- При достижении некоего лимита рабочий набор падает.

Проверка однородности (1)

Проблема

- Составлять отдельную модель для каждой ОС дорого
- Пользовательское версионирование ОС может не отражать версии ядра

Гипотеза

- Предположительно, операционные системы одного поколения одной архитектуры процессора одинаково сконфигурированные имеют схожие алгоритмы работы с памятью

Проверка

- Проверить однородность выборок MS Windows x32 (XP, 7)
- Проверить однородность выборок MS Windows x64 (2008, 7)

Проверка однородности (2)

Используемые критерии

- Критерий Уилкоксона
- Критерий Зигеля-Тьюки
- Критерий Андерсона-Дарлинга

Все критерии показали неоднородность выборок (как оценка объема потребления памяти, так и оценка скорости роста потребления памяти)

Проверка однородности (2)

Используемые критерии

- Критерий Уилкоксона
- Критерий Зигеля-Тьюки
- Критерий Андерсона-Дарлинга

Все критерии показали неоднородность выборок (как оценка объема потребления памяти, так и оценка скорости роста потребления памяти)

Проверка однородности (3)

Однородность значения xp x32, w7 x32

Название критерия	***pValue***	***Однородность***
Wilcoxon	2e-09	Нет
Siegel-Tukey	0	Нет
Anderson–Darling	0	Нет

Однородность дельты w2k8 x64, w7 x64

Название критерия	***pValue***	***Однородность***
Wilcoxon	0	Нет
Siegel-Tukey	0	Нет
Anderson–Darling	0	Нет

Корректировка с учетом обратной связи

- Увеличиваем balloon пока не получим негативной обратной связи. Далее снижаем на 10%_назначенной памяти * ε. ε– жадная модель
- Начальный размер balloon = назначенная память – лимит
- При росте рабочего набора balloon не растет
- Критерий негативной обратной связи:
  - Срабатывание OOM-killer-а на Linux
  - Рост (PageFileSize – PageFileAvailable) на Windows

Анализ: выводы

- Средний уровень рабочего набора для данной системы зависит от многих параметров, в том числе недоступных для метода
- Мы можем предсказать лимит на основании типа ОС и назначенной памяти, но это достаточно сложно
- Мы можем предсказать рост рабочего набора, возникающий на системе под давлением нагрузки
- По обратной связи мы оценим, можно ли уменьшать количество выделенного ресурса памяти без ущерба производительности (= без увеличения страничных промахов)

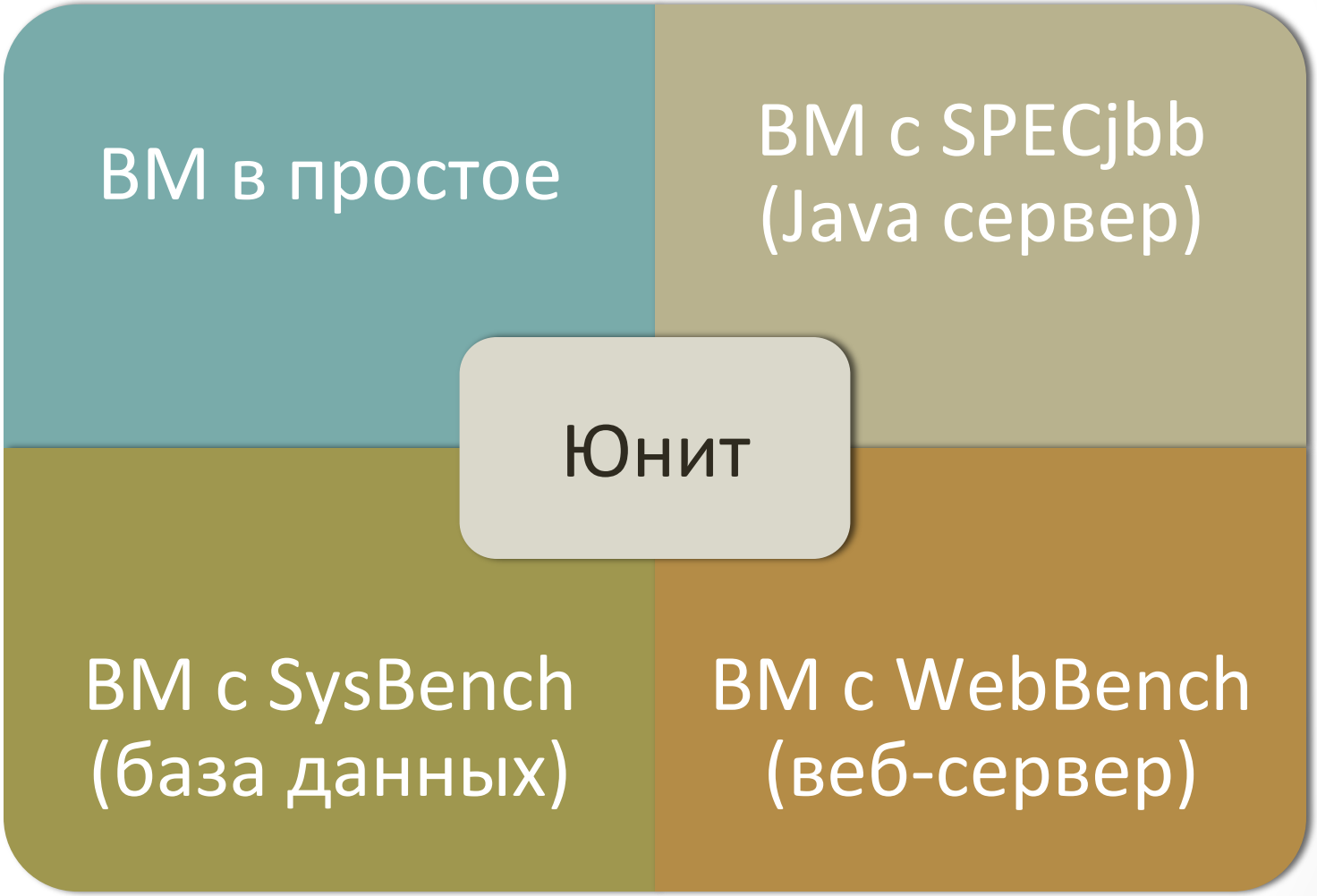
Эмпирическая проверка (1)

Принципы vConsolidate

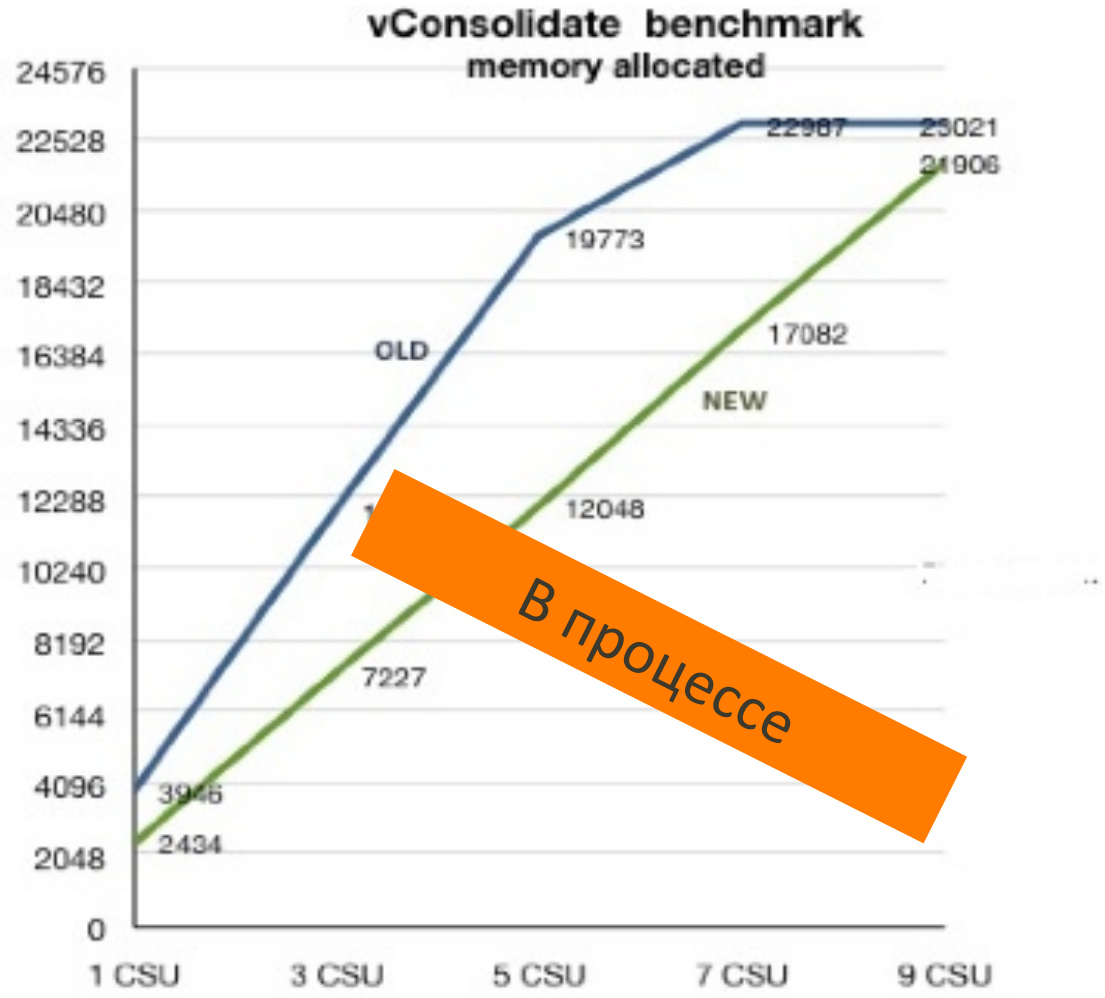
- ВМ в рамках одной физической машины не обязаны иметь однородную нагрузку
- Итоговая оценка теста есть среднее геометрическое оценок от всех тестов

- 1 юнит – набор из ВМ с разными типами тестов
- Соотношение (число юнитов ~ итоговая оценка) отражает эффективность управления памятью
- Соотношение (число юнитов ~ потребленная память) отражает экономию по ресурсам

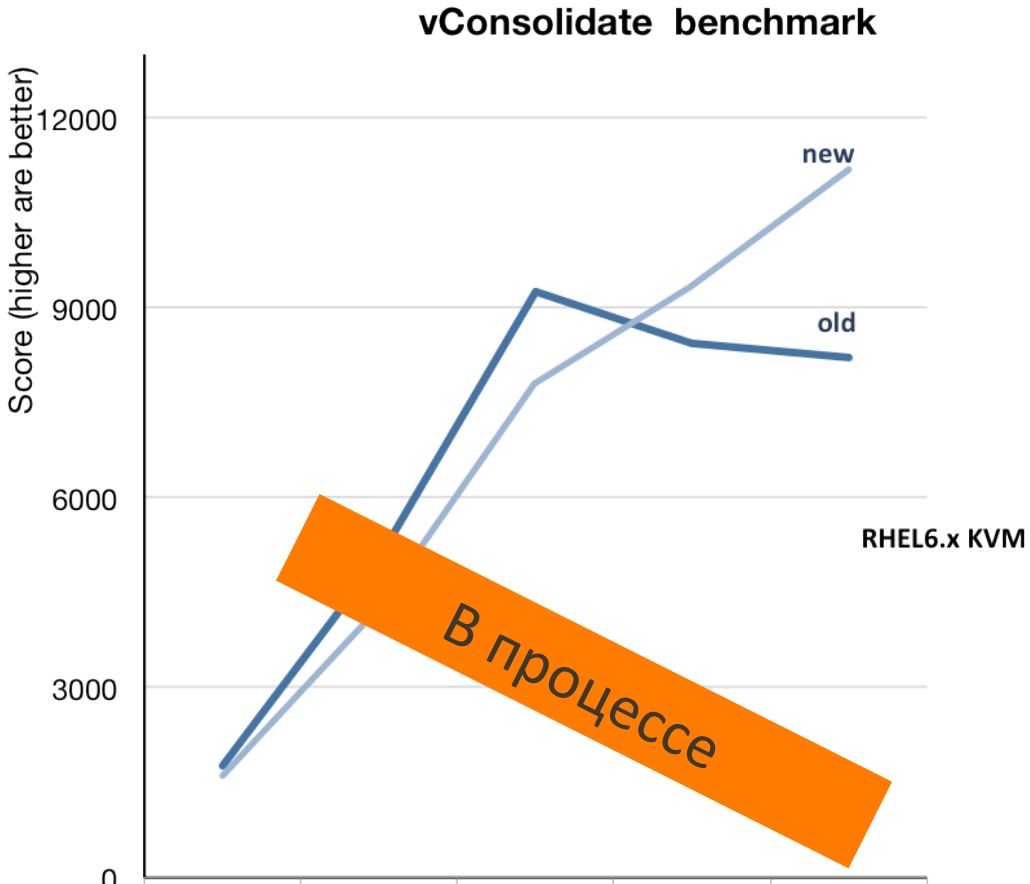
Эмпирическая проверка (2)



Эмпирическая проверка (3). Потребление ресурса



Эмпирическая проверка (3). Производительность



	1 CSU	3 CSU	5 CSU	7 CSU	9 CSU
old	1768	5030	9252	8444	8218
new	1612	4230	7790	9324	11180

Основные результаты

1. Общая концепция предсказания рабочего набора на основании статистики по виртуализационным событиям
2. Способ корректировки результата путем учета внутренних счетчиков гостевой ОС
3. Экспериментальная верификация полученного результата
4. Методы программной реализации предложенных решений, позволяющие в реальном масштабе осуществлять управление ресурсами в распределенной облачной инфрастуктуре.
5. Методы программной реализации по сбору статистики

Спасибо за Ваше внимание

- Предполагается защита по ктн 05.13.11
- Ваши дополнения? Пожелания?

Мелехова Анна

ведущий программист Parallels

anyav@parallels.com

Презентация для
V Международной конференции «Облачные вычисления. Образование. Исследования. Разработка» 2014

Публикации

1. А.Л. Воробьева, «Стратегия виртуализации физической памяти, применяемые в виртуальных машинах». «Процессы и методы обработки информации», Сборник научных трудов, Москва, МФТИ, 2009
2. patent 2354.0250002 “Expansion of virtualized physical memory of virtual machine” pat7925818 http://www.google.com/patents/about?id=R7V6AQAAEBAJ&dq=7925818
3. Воробьева А.Л. «Методика определения размера рабочего набора виртуальной машины», «Математическое моделирование информационных систем», Сборник научных трудов, Москва, МФТИ, 2012
4. Воробьева А.Л., «Управление памятью в гипервизоре. Все о виртуализации памяти в Parallels”, “Разработка высоконагруженных систем», Изд-во Олега Бунина, Москва, 2012
5. Anna Melekhova, «Machine Learning in Virtualization: Estimate a Virtual Machine's Working Set Size», Proceedings on 2013 IEEE Sixth International Conference on Cloud Computing CLOUD 2013
6. А. Бондарь, Д. Карпов, А.Мелехова «Энергосбережение изнутри: что в действительности могут измерить профилировщики», RSDN, 2013 http://rsdn.ru/article/energy_report/EnergyReport_RSDN_3.xml
7. Маркеева Л, Мелехова А, “Проверка гипотезы однородности виртуализационных событий, порожденных различными операционными системами” ”, Труды МФТИ, Москва, 2014, Том 6
8. А. Бондарь, Н. Ефанов, А. Мелехова «Алгоритмы решения задачи динамического управления питанием в облачной системе» (в стадии корректуры для novtex)

Выступления

1. Воробьева А.Л. «Анализ накладных расходов на виртуализацию», Гагаринские чтения 2010, Заседание секции № 26 “Информационные системы и прикладные информационные технологии”
2. Воробьева А.Л. «Опыт создания и развития системы диагностики в виртуализационных продуктах Parallels», SECR 2011
3. Воробьева А.Л. «Управление памятью в гипервизоре. Все о виртуализации памяти в Parallels», Highload 2011
4. Воробьева А.Л. «Трудное наследие разработчиков: сказ о устаревшем коде», SECR 2012
5. Anna Melekhova, «Machine Learning in Virtualization: Estimate a Virtual Machine's Working Set Size», IEEE Sixth International Conference on Cloud Computing CLOUD 2013
6. Anna Melekhova, Alexandr Tormasov, “To migrate or not to migrate: decision based on virtualization patterns”, Proceedings on Second International Conference "Cluster Computing" CC 2013 (Ukraine, Lviv, June 3-5, 2013)