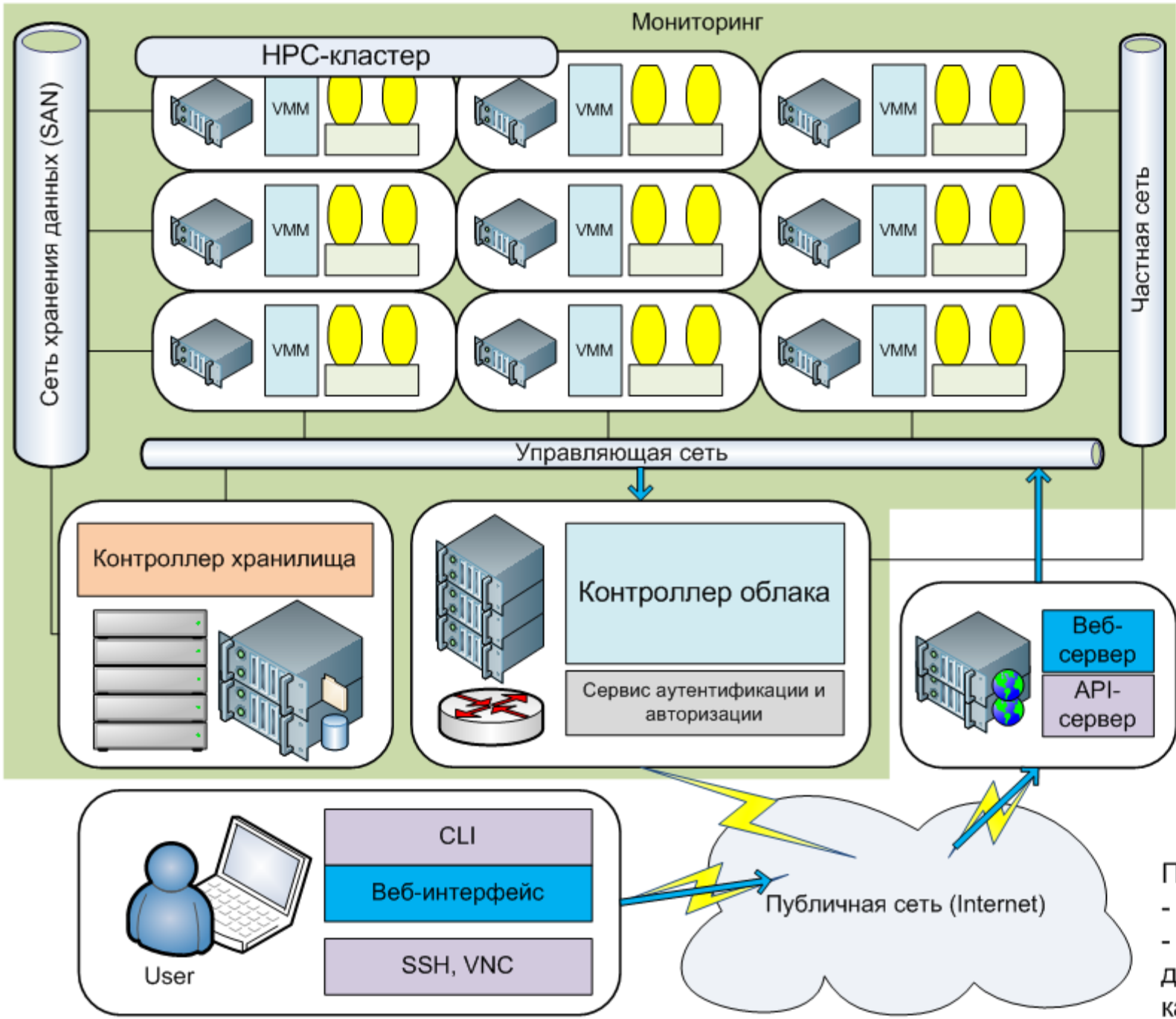
ISP RAS

Перспективы виртуализации высокопроизводительных систем

Кудрявцев Александр Олегович

Институт системного программирования
Российской академии наук

Цель: перенос HPC-вычислений в облако



Облачные платформы:
- Amazon Web Services
- Google Compute Engine
- OpenStack
- CloudStack
- Microsoft Azure

Коммуникационные среды:
- Ethernet
- Infiniband
- Myrinet

Сервис различного уровня:
- Инфраструктура (IaaS)
- Платформа (PaaS)
- Приложение (SaaS)

Инфраструктура:
- Кластер ВМ
- Характеристики оборудования и ПО задаются пользователем

Платформа:
- MPI-кластер как сервис
- Hadoop, OpenFoam и другие HPC-платформы как сервис

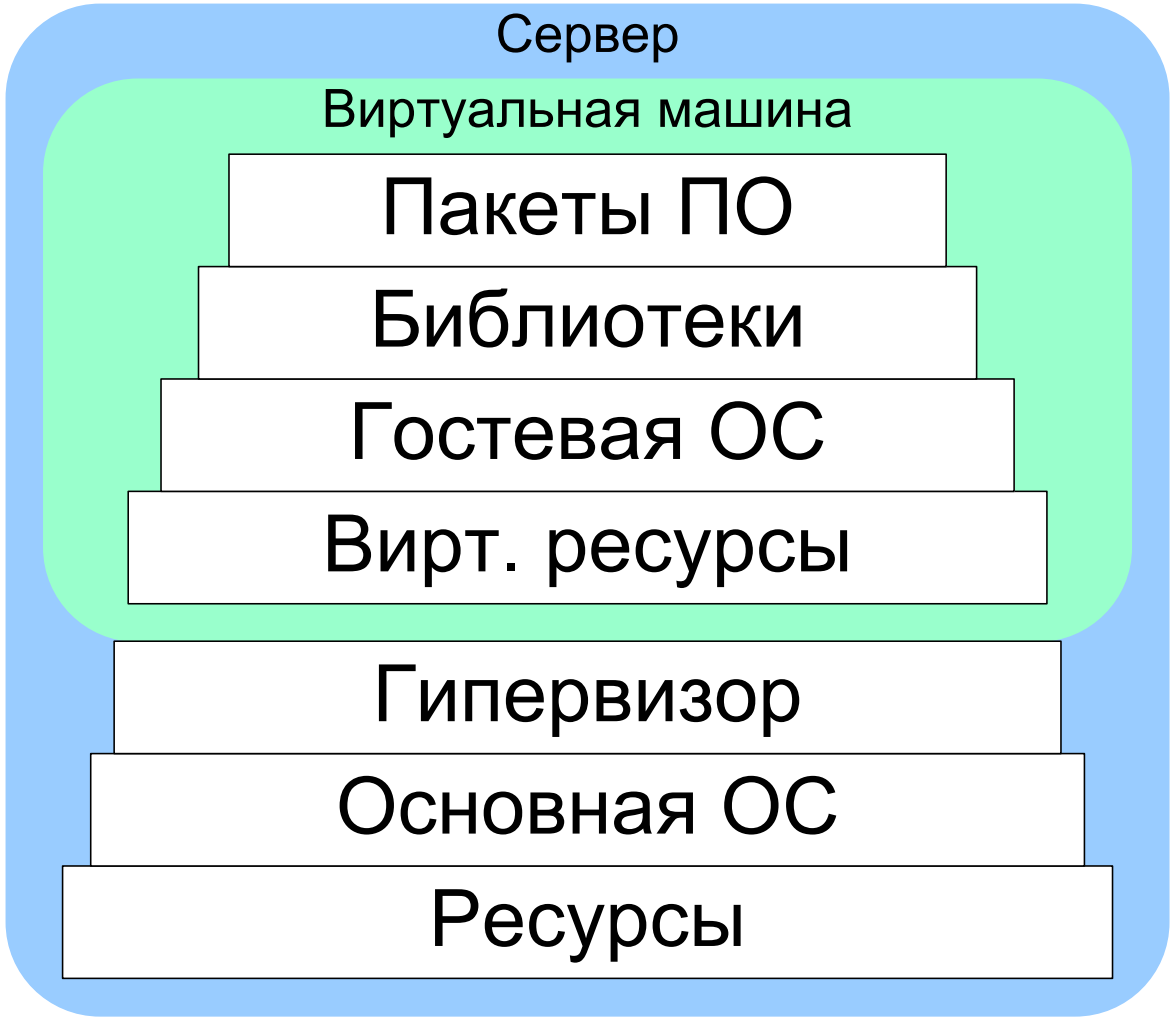
Виртуализация

- Ключевая технология переноса в облако
- Не существует индустриального решения (причина – накладные расходы)
- Несмотря на ограничения, HPC-задачи уже переносятся в облако
  - «Высокомощные» вычисления

Ограничения

- Нельзя полностью избавиться от накладных расходов
  - Виртуализация плохо применима к суперкомпьютерным системам
- HPC-системы (кластеры) среднего масштаба
  - Linux-подобные ОС
  - Открытое ПО
  - **Виртуализация имеет смысл**

Виртуализация: стек ПО



- ПО гостевой системы
- ПО виртуализации

Решаемые задачи

1. Адаптация ПО виртуализации:
   - Обеспечение соответствия виртуальных ресурсов и ресурсов сервера (снижение расходов)
   - Сохранение преимуществ виртуализации
2. Настройка ПО гостевой (тестируемой) системы:
   - Обеспечение эффективной работы HPC-приложений
   - Возможность выявления накладных расходов

Используемое ПО

- ОС GNU/Linux, гипервизор KVM/QEMU
  - Индустриальное решение
  - Проще в использовании чем Xen

- ОС Kitten OS, гипервизор Palacios (V3VEE)
  - Разработан специально для HPC-систем
  - Небольшой объем кода
  - Kitten – легковесное ядро, снижен уровень шума

- Бенчмарки: HPC Challenge, NAS Parallel Benchmarks, SPEC MPI2007

Обзор источников накладных расходов

- Виртуализация процессора
  - накладные расходы близки к нулю
- Виртуализация памяти
  - накладные расходы на преобразование вирт. адреса ВМ в физ. адрес реальной системы
- Виртуализация устройств
  - накладные расходы IOMMU, виртуализации прерываний
- Шум основной ОС

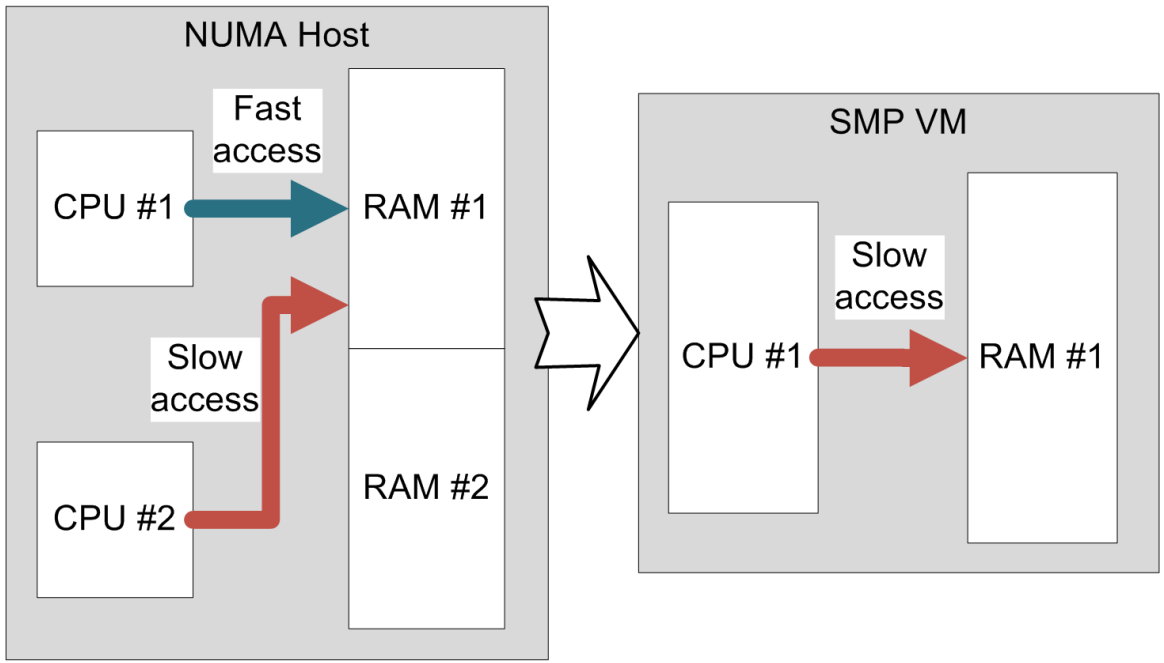
1. Адаптация ПО виртуализации

- Выделение всех ядер процессора ВМ
- Выделение большей части ОЗУ ВМ
- Предоставление ВМ реального коммуникационного устройства

- Привязка выделенных ресурсов к реальным ресурсам
  - Например, серверы архитектуры NUMA

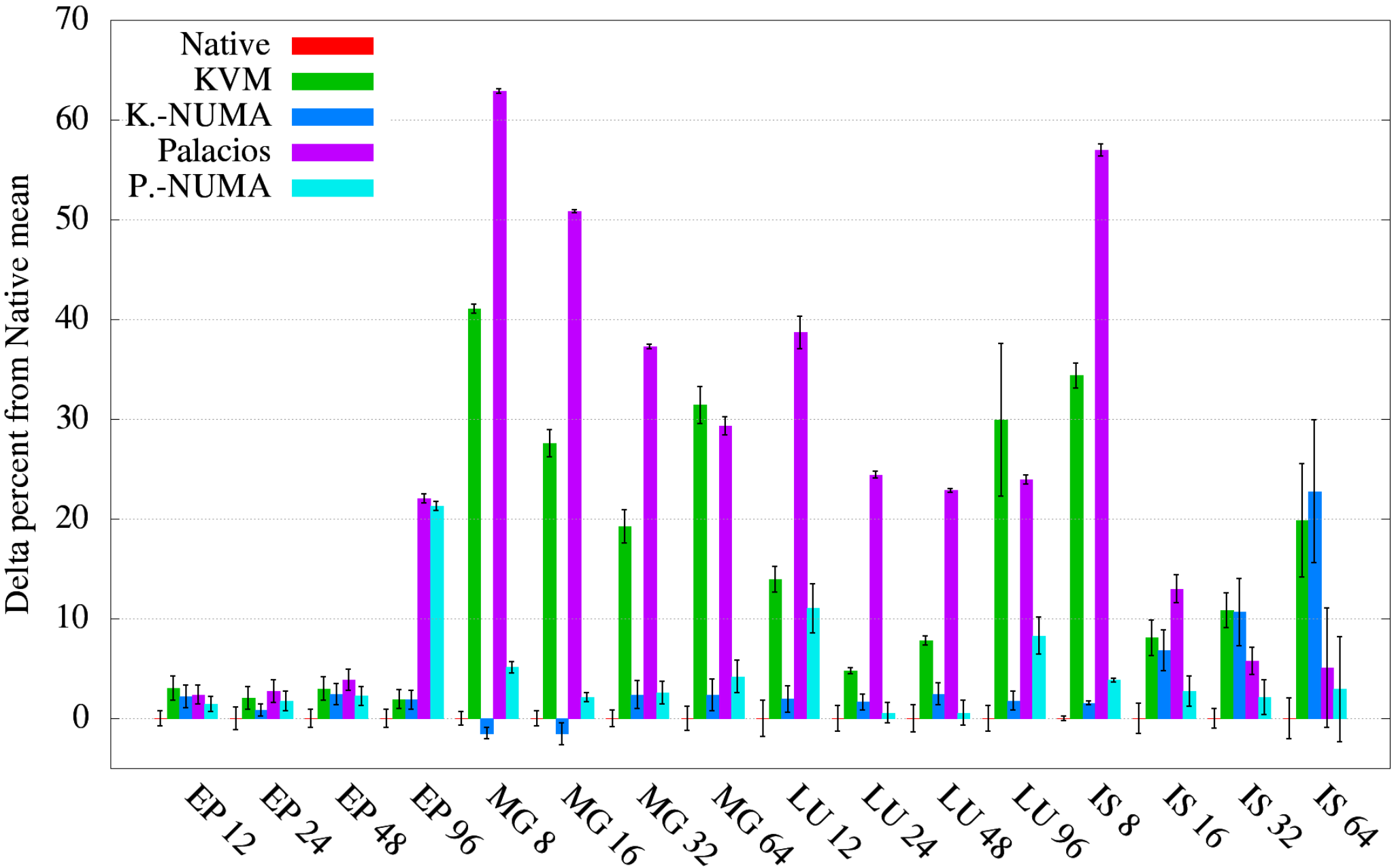
Отражение архитектуры сервера в виртуальной машине

- По умолчанию архитектура ВМ – SMP



- Архитектура многих современных серверов – NUMA

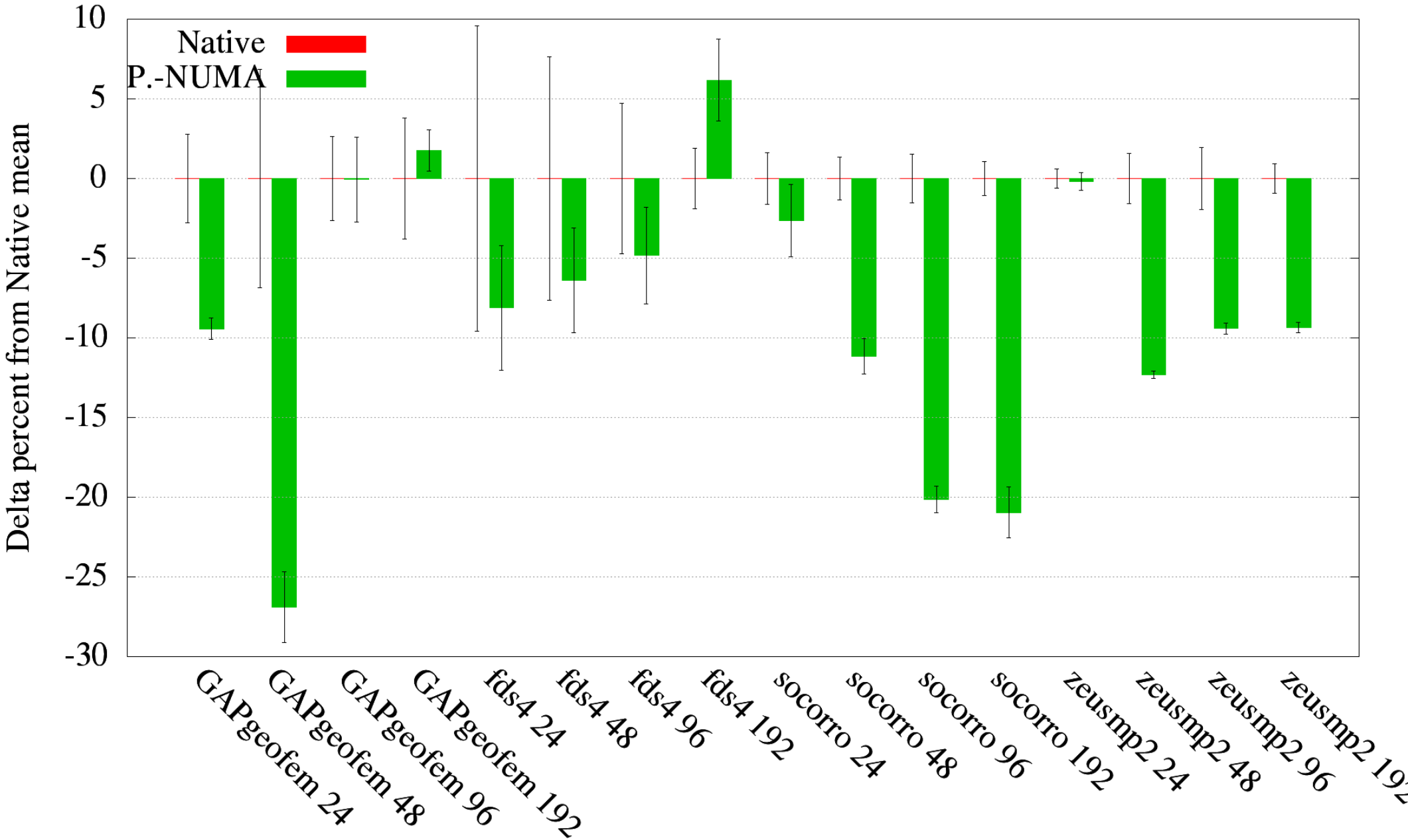
Влияние эмуляции NUMA (NAS Parallel Benchmarks)



2. Настройка ПО тестируемой системы

- Неоптимальная настройка может привести к невозможности выявления накладных расходов виртуализации
  - Например, накладные расходы маскируются простаивающим процессором
  - В том числе возможно улучшение производительности программы в ВМ по сравнению с реальным сервером

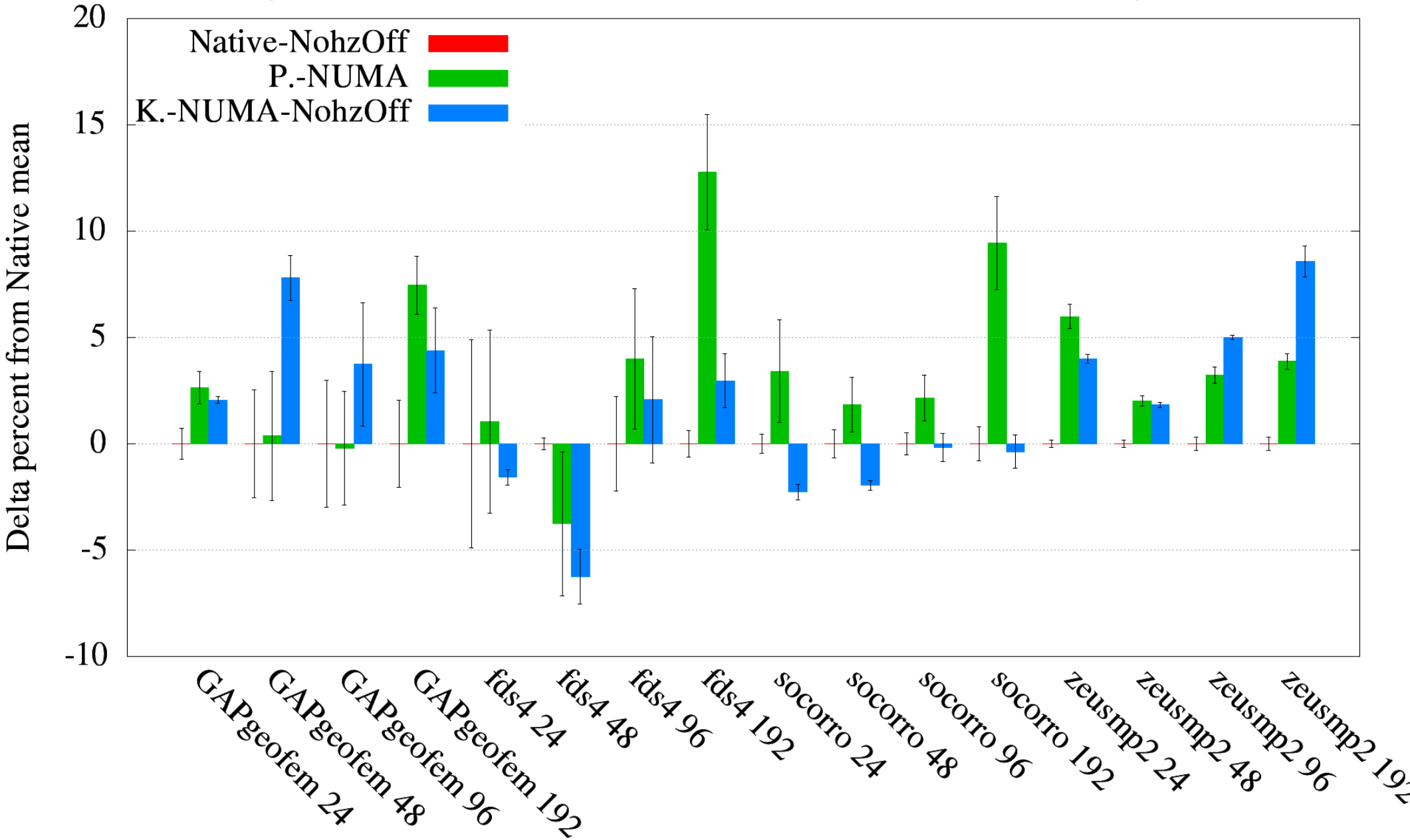
Улучшение производительности в BM (Palacios, SPEC MPI2007)



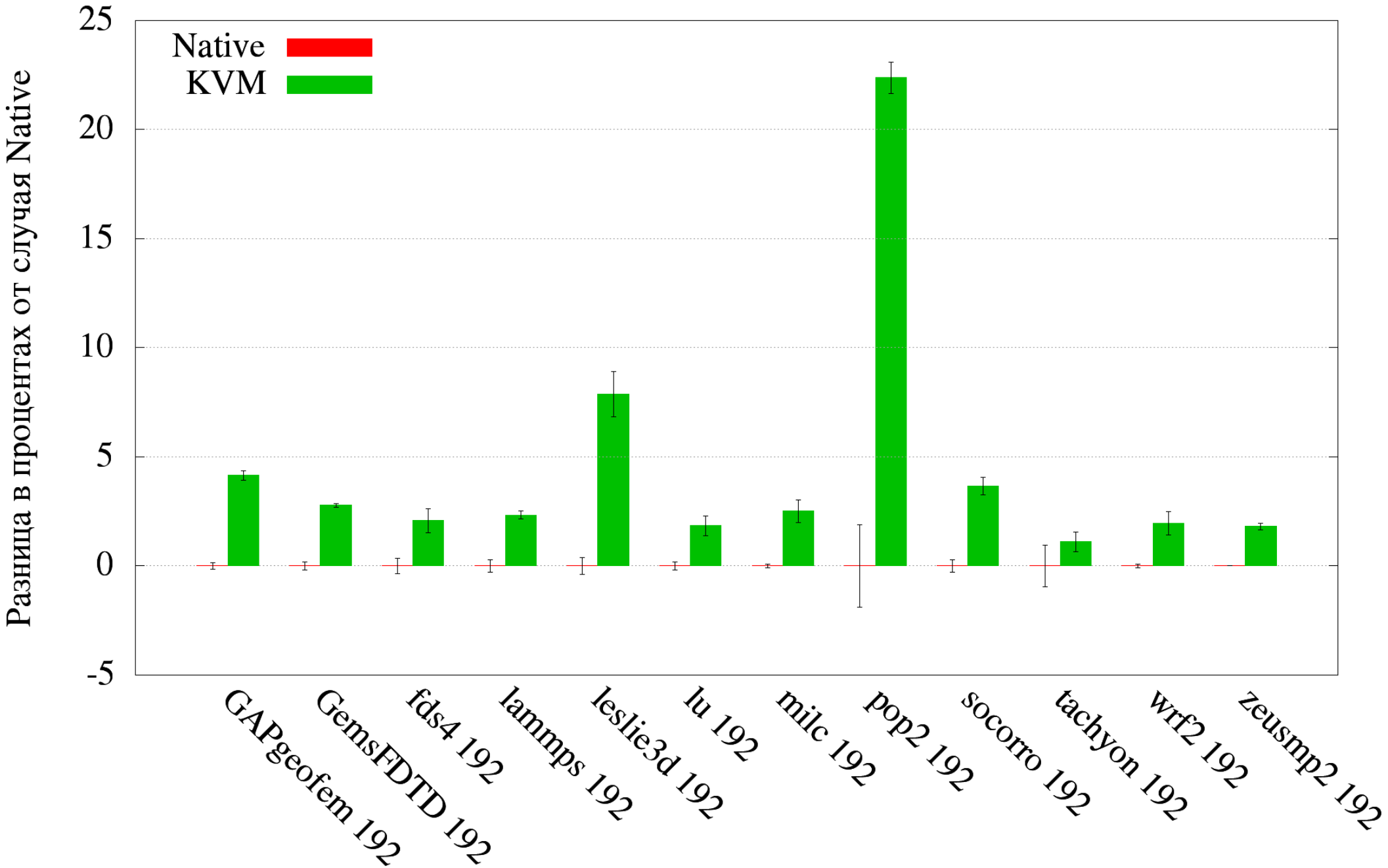
Улучшение производительности в ВМ

- Профилирование показало замедление функции MPI_Waitall в случае реального сервера
- Причина: неоптимальная конфигурация ПО тестовой системы
  - Несогласованность библиотеки OpenMPI и драйвера Infiniband приводила к частым переходам процессора в энергосберегающий режим с отключением таймера (механизм NoHz)
  - Как следствие, резко увеличивается задержка
  - В ВМ Palacios механизм отключения таймера не работает из-за недостатков виртуального окружения

Отключение механизма NoHz (KVM, Palacios, SPEC MPI2007)



Результаты SPEC MPI 2007 с использованием библиотеки MVAPICH

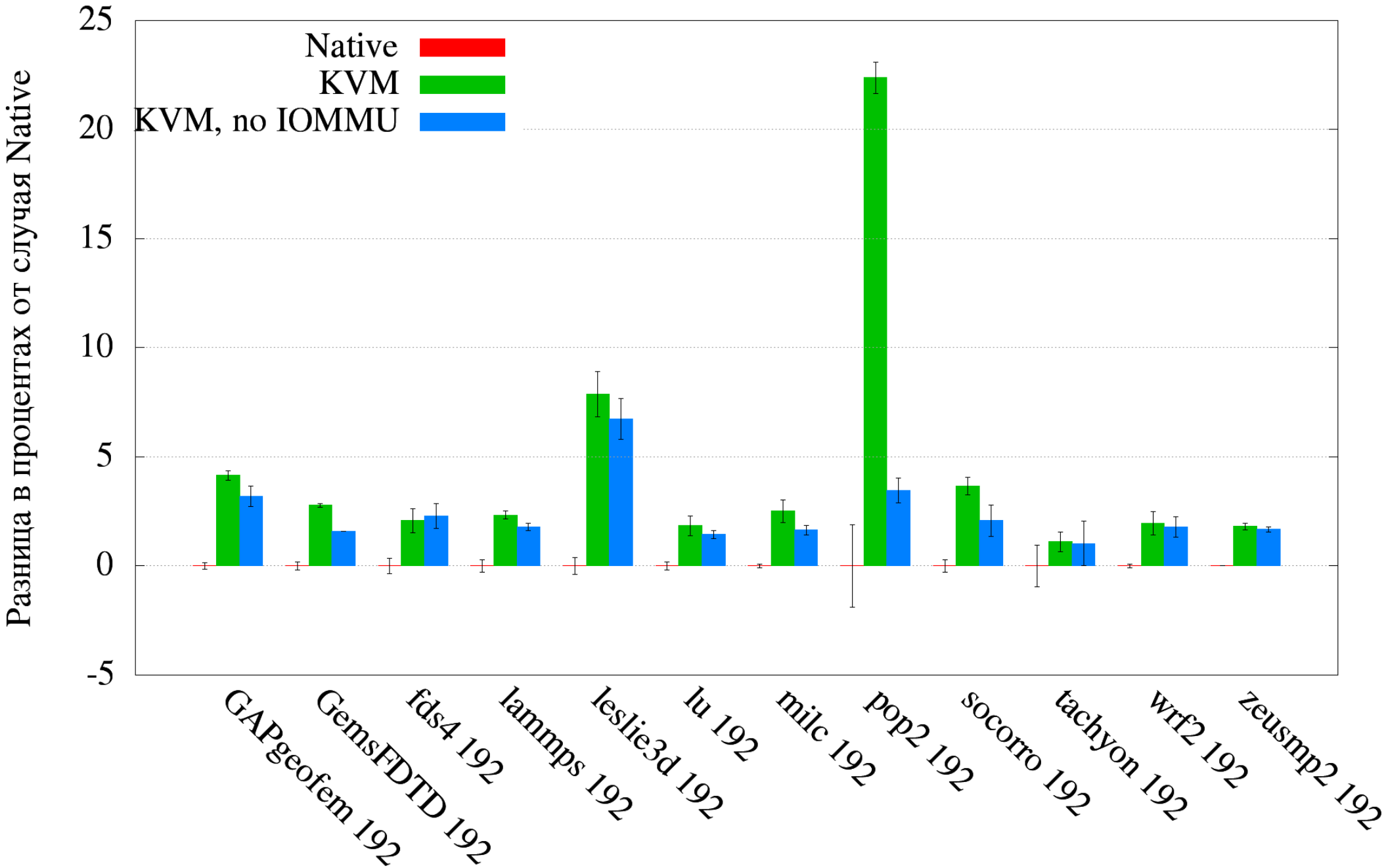


Производительность теста pop2

- Причина накладных расходов – использование устройства виртуализации ввода-вывода (IOMMU)
- IOMMU транслирует адреса устройства в физические адреса памяти ВМ при выполнении DMA-транзакций
- Узкое место – кеш трансляции адресов
- Временное решение: обход IOMMU с использованием паравиртуального интерфейса

Результаты SPEC MPI 2007 с обходом IOMMU

ISP RAS



Заключение

- Эффективная виртуализация возможна для широкого класса HPC-приложений
  - Накладные расходы не более 7%
  - Система виртуализации KVM/QEMU позволяет достичь высокой производительности BM
- Необходима тщательная доводка системы:
  - Адаптация ПО виртуализации, соответствие виртуальных ресурсов реальным ресурсам сервера
  - Настройка ПО тестовой системы для повышения производительности

Планируемые работы

- Увеличение масштаба экспериментов (1024 ядра)
- Интеграция доработанной системы виртуализации в облачную платформу OpenStack
- Исследование приложений, показавших наибольшее падение производительности (leslie3d)
- Тестирование прикладных пакетов (OpenFOAM)
- Использование различного оборудования (Ethernet)

Спасибо!