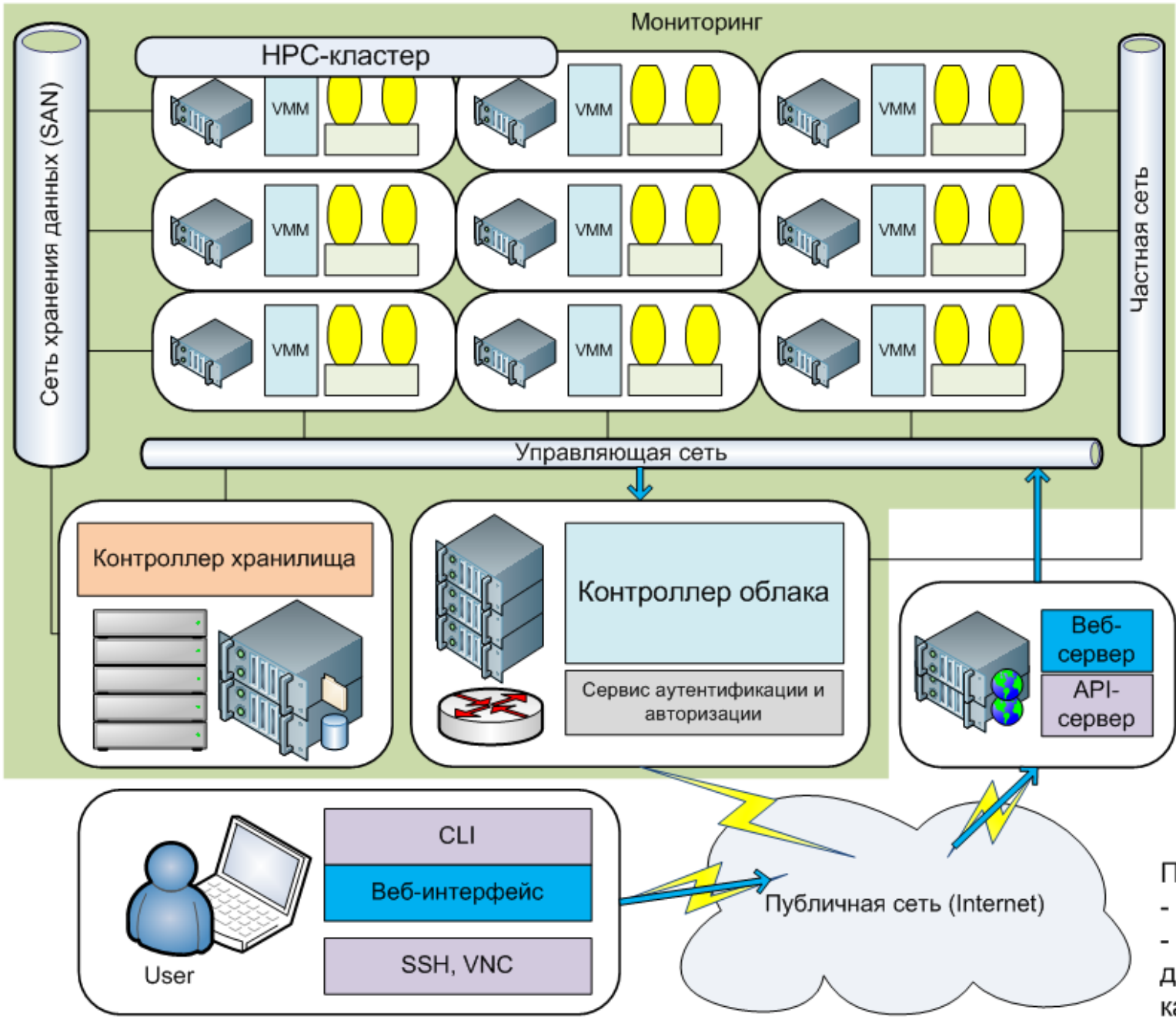
ISP RAS

Высокопроизводительные вычисления как облачный сервис: ключевые проблемы

Кудрявцев Александр Олегович
(alexk@ispras.ru)

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

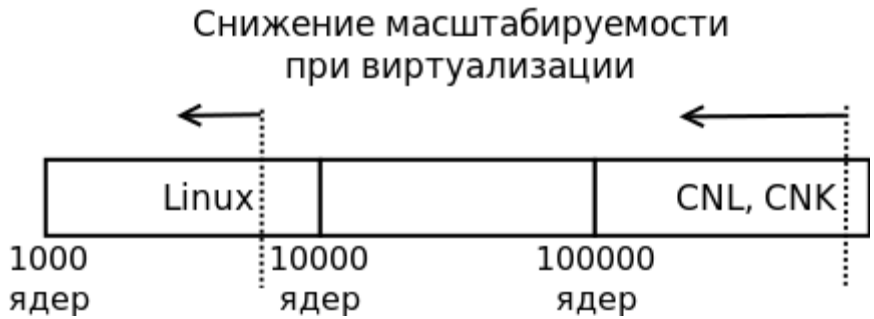
Проблема – переносимость производительности, масштабируемость

Два варианта реализации HPC-облака

- Виртуализация платформы (CPU, память, устройства)
  - Гибкость, безопасность
  - Широкая область применения
  - Накладные расходы (коммуникационная среда, шум ОС)
- Контроль на уровне оборудования (“bare-metal”)
  - Нет снижения производительности
  - Запуск гостевой ОС без контроля гипервизора
  - Требуется специализированное оборудование, тщательная настройка

Ограничения на применение виртуализации

- Нельзя полностью избавиться от накладных расходов
  - Виртуализация неприменима к суперкомпьютерным системам



- **Задача:**
  - Накладные расходы не более 10% при масштабируемости не менее чем до 1024 ядер

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
- Промышленный пакет OpenFOAM

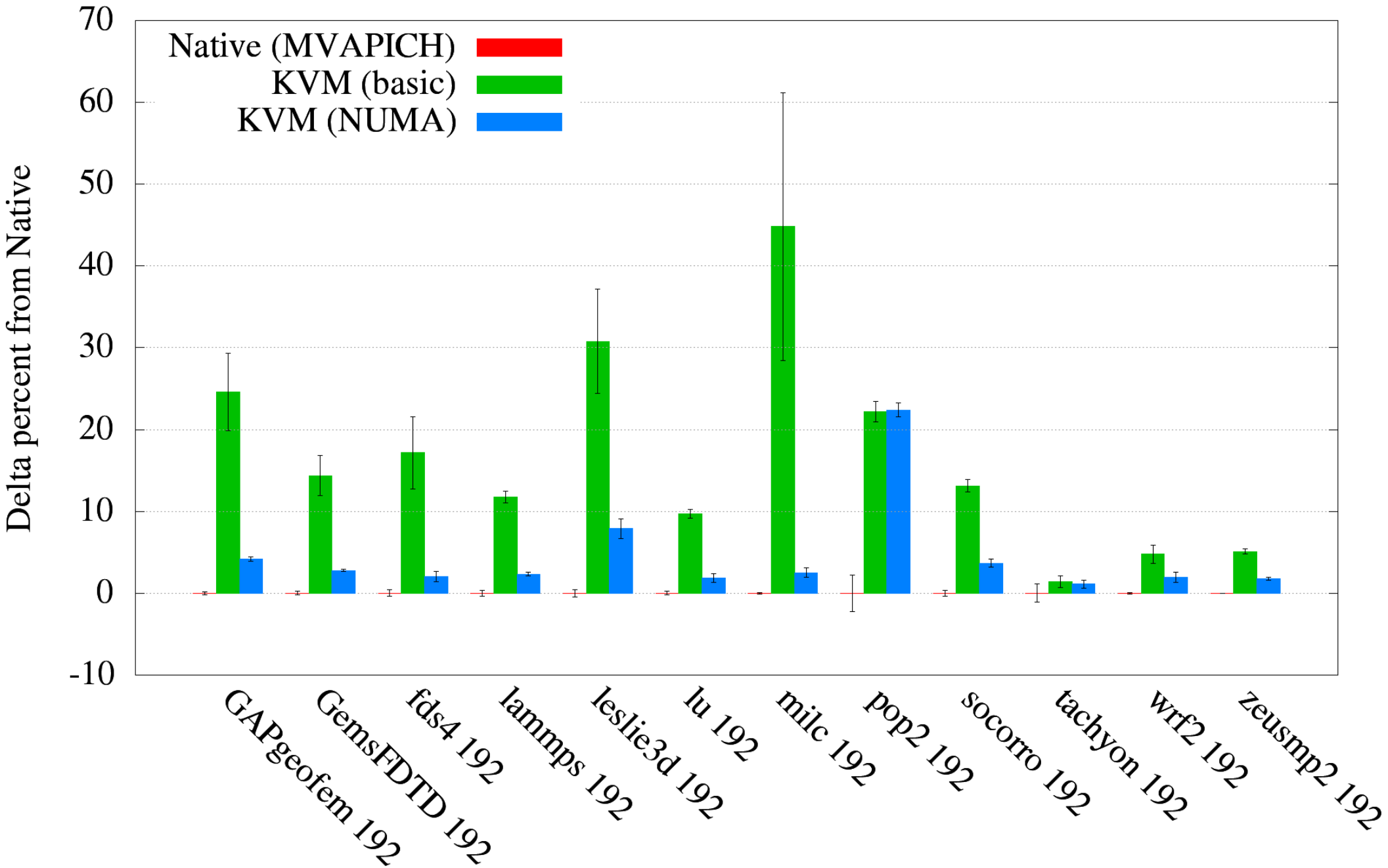
Источники накладных расходов при виртуализации

- Виртуализация памяти
  - накладные расходы на преобразование вирт. адреса ВМ в физ. адрес реальной системы
- Виртуализация устройств
  - необходим проброс устройств:
    - накладные расходы IOMMU
    - виртуализация прерываний
- Шум основной ОС
  - снижение масштабируемости производительности

Способы сокращения накладных расходов для HPC-задач

- Выделение всех ядер процессора ВМ
- Выделение большей части ОЗУ ВМ
  - С использованием больших страниц и вложенной страничной адресации
- Привязка выделенных ресурсов к реальным ресурсам
  - Серверы архитектуры NUMA => ВМ архитектуры NUMA
- Предоставление ВМ реального коммуникационного устройства
  - В том числе в обход IOMMU
- Снижение частоты таймера ОС с 1000 до 100 ГЦ
  - Позволяет снизить шум

Влияние привязки NUMA, SPEC MPI2007

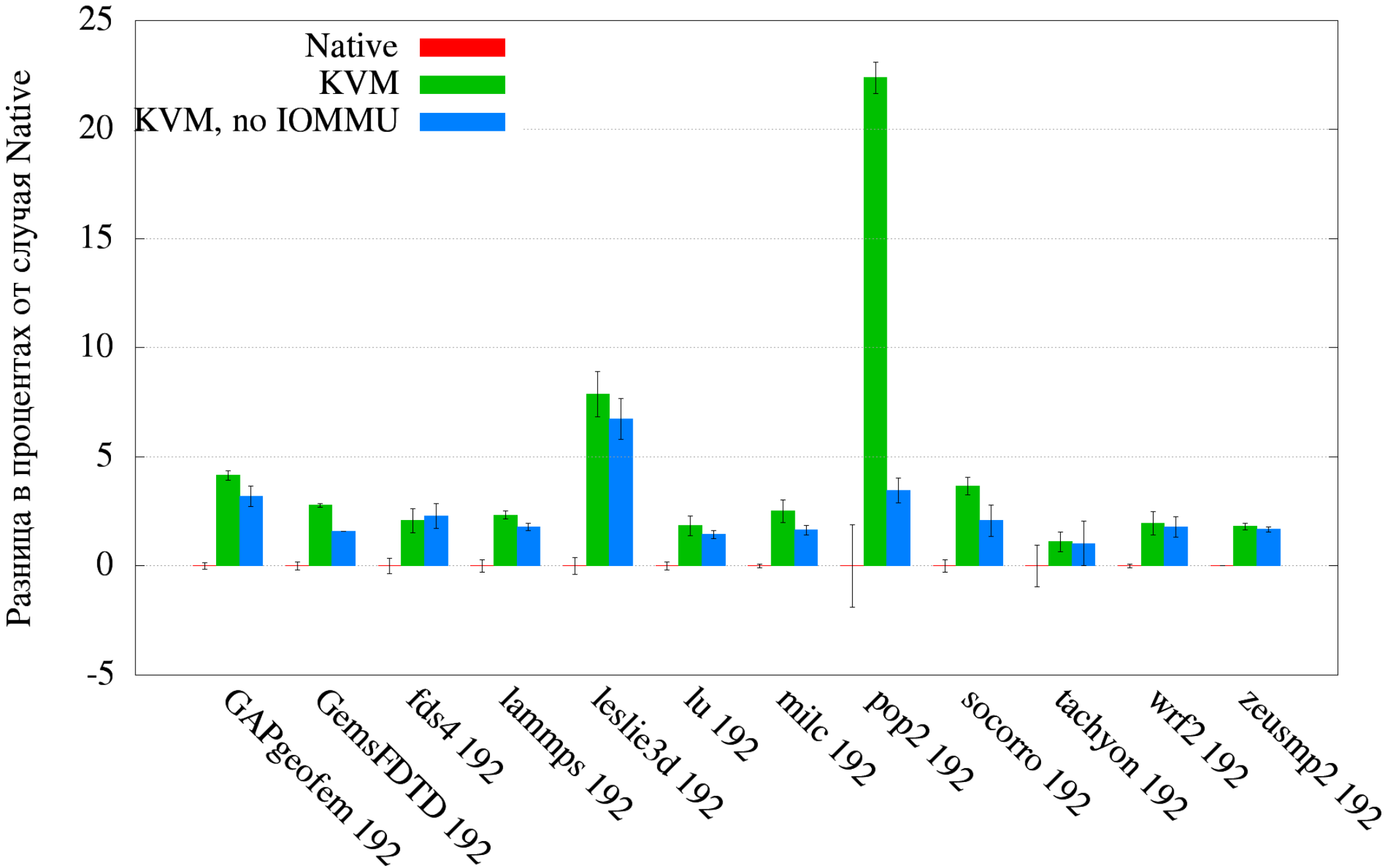


Производительность теста pop2

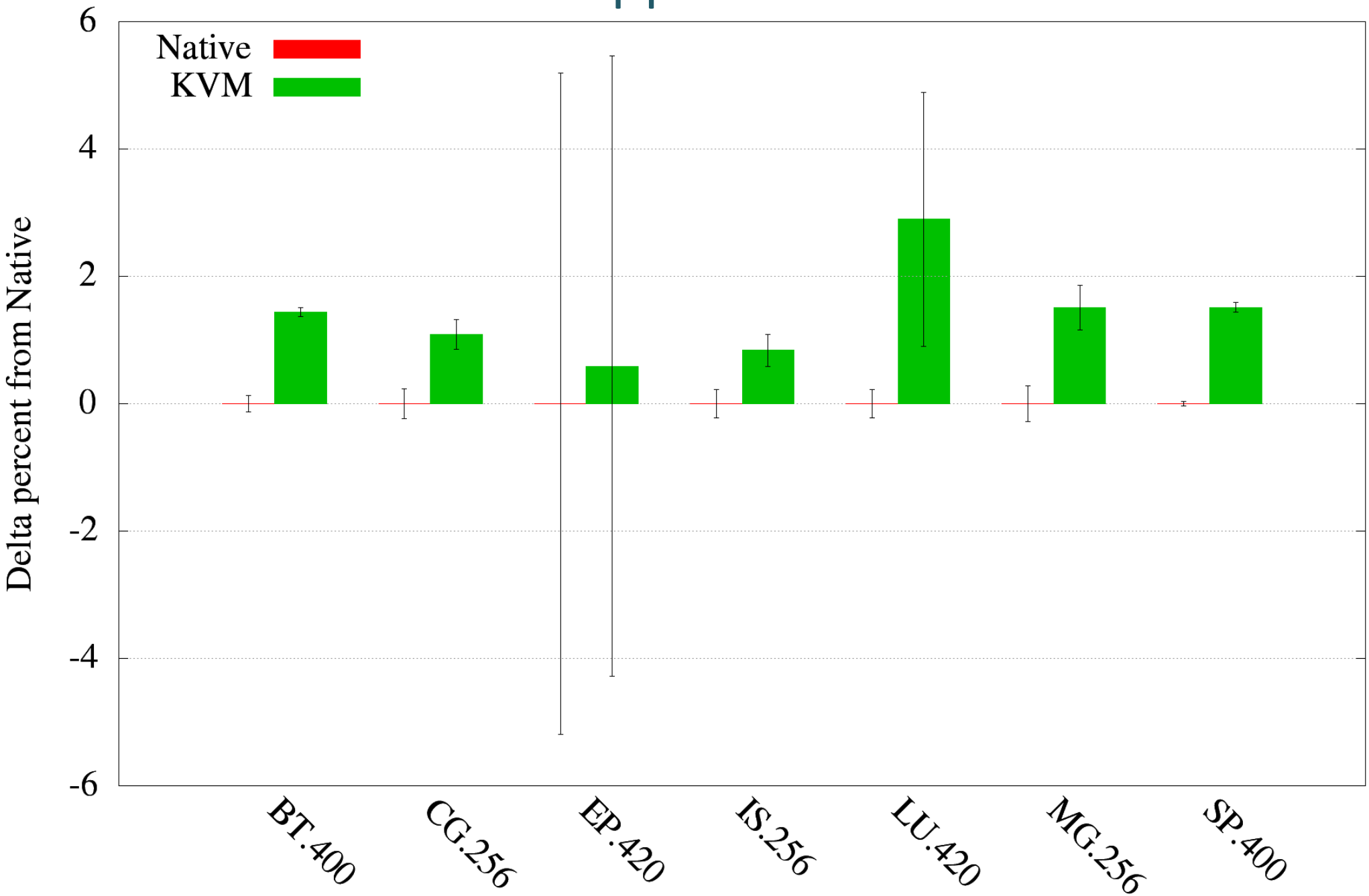
- Причина накладных расходов – использование устройства виртуализации ввода-вывода (IOMMU)
- IOMMU транслирует адреса устройства в физические адреса памяти ВМ при выполнении DMA-транзакций
- Узкое место – кеш трансляции адресов
- Временное решение: обход IOMMU с использованием паравиртуального интерфейса

Результаты SPEC MPI 2007 с обходом IOMMU

ISP RAS



Результаты NAS Parallel Benchmarks с обходом IOMMU



Заключение

- Эффективный перенос в облако возможен для достаточно широкого класса HPC-приложений
  - Накладные расходы не более 10%
  - Запуск до 420 процессов
  - Система виртуализации KVM/QEMU позволяет достичь высокой производительности BM, сохраняя достаточный уровень гибкости
  - Небезопасный проброс коммуникационного адаптера в BM
- В ряде случаев, достигнуть достаточной производительности не удается:
  - “Мелкозернистые” групповые коммуникации
  - Кластеры на базе Ethernet

Текущие и планируемые работы

- Интеграция разработанной системы виртуализации в облачную инфраструктуру программы “Университетский кластер”
- Тестирование производительности на большем числе ядер (не менее 1024)
- Тестирование прикладных пакетов (OpenFOAM)

- Исследование вариантов применения концепции “bare-metal” облаков для HPC-задач