Отчёт по работе с кластером ИВЦ НГУ

- Тема работы: Адаптация теста HPL для оценки производительности неоднородных вычислительных систем.
- Состав коллектива:
  Ефимцев Фёдор Андреевич - выполнял работу;
  Городничев Максм Александрович - куратор работы, ст. преп. каф. ПВ ФИТ НГУ, асс. каф. ПВТ НГТУ.
- Информация о гранте: нет
- Постановка задачи:
  Исследовать модификацию теста-бенчмарка HPL, поддерживающую отгрузку части вычислений на GPU, на вычислительных системах-“зоопарках” (состоящих из отличающихся по набору железа узлов). Предложить алгоритм подбора оптимальных параметров теста (т.е. таких, при которых достигается максимальая производительность) в зависимости от исходных параметров вычислительных узлов системы.
- Актуальность проблемы заключалась в необходимости изучения принципов работы (т. е. выполнения произвольных задач) с неоднородными вычислительными системами, включающими в себя CPU и GPU. В качестве примера такой задачи была взята модификация теста HPL, HPL-GPU (https://github.com/davidrohr/hpl-gpu/), в которой добавлена отгрузка части вычислительных операций (например, матрично-матричное умножение) на GPU при помощи API CUDA или OpenCL.
- Ход работы:
  На первом этапе производились анализ исходного кода HPL-GPU, изучение его документации, сборка и тестовые запуски программы теста. Было необходимо найти на кластере (или скачать) и подключить все необходимые библиотеки (Intel MKL, TBB, CUDA, CALDGEMM), с чем периодчески возникали сложности из-за несовместимости конкретных версий. Тестовые запуски показали, что программа бенчмарка выдаёт близкие к ожидаемым результаты при использовании только CPU, но при выполнении на очереди **a6500g10** с использованием одного ускорителя NVIDIA Tesla V100 результат бенчмарка получался всего лишь на уровне 200-300 Гфлопс, что более чем на порядок ниже ожидаемых значений.
  На следующем этапе предполагалось оптимизировать работу теста – выяснить причину медлительной его работы и на основе документации к бенчмарку подобрать значения для оптимальной производительности. Затем предпологалось умышенно “замедлять” различные компоненты вычислительной системы (частоту процессора, количество используемых CUDA-ядер и так далее), или же напротив увеличить количество используемых узлов и проанализировать для полученных систем результаты, при необходимости также подбирая параметры бенчмарка. Данный этап не был завершён.
- Эффект от использования кластера в достижении целей работы: кластер ИВЦ НГУ дал возможность поработать с реальной высокопроизводительной вычислительной системой, в составе которой присутствуют современные GPU. Существенно облегчил тестирование и настройку программы бенчмарка тот факт, что очередь **a6500g10_short** на момент выполнения работы была практически всегда свободна.