**Высокопроизводительные вычисления на кластерах с использованием графических ускорителей NVIDIA**

| Лекция | Практикум |
|---|---|

| 1 июля. Вводная часть. |
|---|
| Системы с графическими ускорителями в Российских учебно-научных центрах. |
| Принципы работы графических ускорителей, программная модель CUDA. |
| Вычислительный комплекс с GPU: получение информации о системе, средства мониторинга и диагностики, среды разработки и исполнения приложений. |
| Устройство CUDA-компилятора: стадии обработки кода, промежуточные представления, загрузка ядер, JIT-компиляция. |

| 2 июля. Эффективные алгоритмы и быстрая разработка. |
|---|
| Иерархия памяти CUDA, эффективное использование разделяемой памяти. Общее виртуальное адресное пространство (UVA). |
| Разработка CUDA-приложений на языке Fortran. ISO_C_BINDING, замечания о способе передачи аргументов. |
| Эффективная реализация некоторых алгоритмов с использованием разделяемой памяти и UVA. |
| Быстрая разработка CUDA-приложений на C++ с помощью Thrust. Библиотека алгоритмов линейной алгебры с разреженными матрицами CUSP. Использование в C и Fortran. |
| Реализация некоторых алгоритмов с помощью Thrust. |

| 3 июля. Прикладные библиотеки. |
|---|
| Прикладные библиотеки со встроенной поддержкой GPU, часть I: CUBLAS, MAGMA, CUSPARSE, CUFFT, CURAND. |
| Практикум |
| Практикум |

| |
|---|
| Прикладные библиотеки со встроенной поддержкой GPU, часть II: PetSc, Trilinos. |
| Практикум |

| 4 июля. MultiGPU. |
|---|
| Асинхронное исполнение, CUDA Streams. Измерение времени, CUDA Events. Управление несколькими GPU: взаимодействие CUDA с другими программными моделями параллельных вычислений. |
| Реализация конкуретного исполнения нескольких ядер на GPU с промежуточными синхронизациями. Асинхронные и блокирующие операции. |
| Параллельное использование нескольких GPU в последовательном приложении. Несколько GPU в многопоточном приложении на основе интерфейса POSIX. |
| Расширения OpenMPI для CUDA, обмен данными в памяти GPU с помощью MPI |
| Реализация взаимодействия между несколькими CUDA-приложениями с помощью интерфейса IPC |

| 5 июля. Анализ, диагностика, отладка. |
|---|
| Средства анализа, диагностики и отладки CUDA-приложений. Профилировка с помощью CUDA Profiler, диагностика ошибок памяти (cudamemcheck), интерактивная отладка GPU-ядер (cuda-gdb). |
| Анализ эффективности приложения с помощью CUDA Profiler. Основные аппаратные счетчики. |
| Демонстрация работы отладчика: основные возможности, стандартные сценарии использования. Типичные ошибки в приложениях. |
| Программируемая профилировка с помощью CUPTI. |

| 6 июля. Оптимизация программ, архитектура и внутренее устройство GPU. |
|---|
| Архитектура GPU, методы анализа и оптимизации приложений. |
| |

| |
|---|
| Управление кэшем GPU, эффект на производительность при различных шаблонах доступа к памяти. |
| Язык промежуточного представления программы PTX и Fermi ISA. Формат исполняемого образа ядра CUBIN, начальная загрузка. Ассемблер и дизассемблер. |
| Анализ эффективности компилятора на низком уровне: распределение регистров, локальная память, векторизация. Отладка GPU-программы без исходного кода. |