РОСАТОМ

РФЯЦ ВНИИЭФ

ГОСУДАРСТВЕННАЯ КОРПОРАЦИЯ ПО АТОМНОЙ ЭНЕРГИИ «РОСАТОМ»

Тестирование многопроцессорных Супер-ЭВМ с гетерогенной и гибридной архитектурой

Докладчик: А. В. Алексеев.
Соавторы: Н. Р. Антипина, А. В. Ветчинников, А. Н. Залялов, Д. И. Липов, А. В. Ломтев, А. А. Нуждин, И. С. Сапронов

НСКФ-2013
Переславль-Залесский
26 ноября 2013 г.

Цели тестирования

- Определение общей производительности мультипроцессорной вычислительной системы (МВС)
- Исследование характеристик отдельных подсистем и компонент
- Оценка производительности и эффективности распараллеливания на характерных классах задач
- Оценка работоспособности и надёжности МВС
- Подтверждение соответствия характеристик МВС заявленным параметрам
- Сравнение характеристик МВС с аналогичными системами
- Анализ характеристик МВС в ходе разработки (co-design)

Набор тестов ВНИИЭФ

Специальные тесты

b_eff	(Effective Bandwidth Benchmark) производительность коммуникационной среды
IMB	(Intel MPI Benchmark) производительность коммуникационной среды
Stream	производительность памяти вычислительного узла
HPL	производительность вычислительного узла и системы в целом
sPPm	производительность системы хранения

Международные прикладные тесты

NAS	производительность мультипроцессорных систем на примере различных алгоритмов (EP, MG, CG, FT, IS, LU, SP, BT)
sPPM	производительность МВС и эффективности распараллеливания задачи решения 3D уравнений газовой динамики Эйлера

Методические прикладные тесты (разработка ВНИИЭФ)

TDU	эффективность распараллеливания при решении 3D уравнения диффузии нейтронов
GD2	эффективность распараллеливания при решении 3D газовой динамики
PAUK	эффективность распараллеливания при решении 3D уравнения переноса нейтронов DSn-методом
C-MK	эффективность распараллеливания при решении 3D уравнения переноса нейтронов методом Монте-Карло
Egida-Test	эффективность распараллеливания при решении 3D уравнений газовой динамики и теплопроводности
MoDyS	эффективность распараллеливания при решении 3D уравнений молекулярной динамики

Тестовая программа TDU

Год создания:	**1995**
Физическая постановка:	**Перенос нейтронов в трёхмерном многогрупповом диффузионном приближении.**
Численные методы:	**Неполное разложение Холецкого, метод сопряжённых градиентов.**
Распараллеливание:	**1D декомпозиция по третьему пространственному направлению. MPI.**
Аппаратная платформа:	**Универсальные процессоры, Intel Xeon Phi.**
Язык:	**Фортран 77.**
Особенности:	**Интенсивная работа с оперативной памятью.**
Параметры теста:	**Размер пространственной сетки. Число итераций 3-х уровней, определяющих соотношение между вычислительной работой и количеством обменов данными.**
Развитие:	**3D декомпозиция, OpenMP.**

Тестовая программа ГД2

Год создания:	**1999**
Физическая постановка:	**Расчёт движений сплошной среды путём решения системы многомерных уравнений газовой динамики в лагранжево-эйлеровой постановке.**
Численные методы:	**Неявная разностная схема с расщеплением по пространственным направлениям. Решение СЛАУ с трёхдиагональной матрицей методом прогонок.**
Распараллеливание:	**2D декомпозиция по двум пространственным (эйлеровым) направлениям. MPI. Работа с гетерогенными системами. Массово-параллельно-конвейерное распараллеливание решения потоков прогонок.**
Аппаратная платформа:	**Универсальные процессоры.**
Язык:	**Фортран 90.**
Особенности:	**Интенсивная работа с оперативной памятью. Экономичные численные методы.**
Параметры теста:	**Размер пространственной сетки, «вес» уравнения состояния газа, число порций прогонки.**
Развитие:	**3D декомпозиция, OpenMP, распараллеливание с использованием GP-GPU.**

Тестовая программа ПАУК

Год создания:	**2006**
Физическая постановка:	**Стационарный перенос нейтронов в трёхмерном одногрупповом кинетическом приближении.**
Численные методы:	**Схема типа DSn-метода, метод простых итераций, алгоритм бегущего счета.**
Распараллеливание:	**3D пространственная декомпозиция. Метод увеличения. MPI. Алгоритм конвейерного типа (аналог KBA-алгоритма). Аналитическая формула эффективности. Работа с гетерогенными системами.**
Аппаратная платформа:	**Универсальные процессоры, Intel Xeon Phi.**
Язык:	**Фортран 90.**
Особенности:	**Интенсивные двухточечные межпроцессорные обмены с «соседями».**
Параметры теста:	**Размер пространственной сетки. Число направлений полета частиц. Число рассчитываемых слоев между обменами. Параметры пространственной декомпозиции. Число итераций.**
Развитие:	**Смешанное распараллеливание MPI+OpenMP, векторизация вычислений.**

Тестовая программа С-МК

Год создания:	**Вариант 1: 1995 (С-95,С-МК),** **Вариант 2: 2011 (СМК-У версия для GPU)**
Физическая постановка:	**Перенос нейтронов в трёхмерном приближении.**
Численные методы:	**Метод Монте-Карло**
Распараллеливание:	**С-МК: MPI, асинхронное, по траекториям частиц.** **СМК-У : Векторное распараллеливание для арифметических ускорителей**
Аппаратная платформа+ Язык :	**С-МК: Универсальные процессоры, Фортран 90** **СМК-У : GPU (NVIDIA), C++ и CUDA**
Особенности:	**Нет программных ограничений на число процессоров.** **На гетерогенном поле могут использоваться одновременно разные процессоры и арифметические ускорители.**
Параметры теста:	**Заказные результаты – работа с памятью,** **Время записи результатов – нагрузка на файловую систему** **Время межпроцессорного обмена – нагрузка на коммуникации**
Развитие:	**Арифметические ускорители (CUDA), OpenMP.**

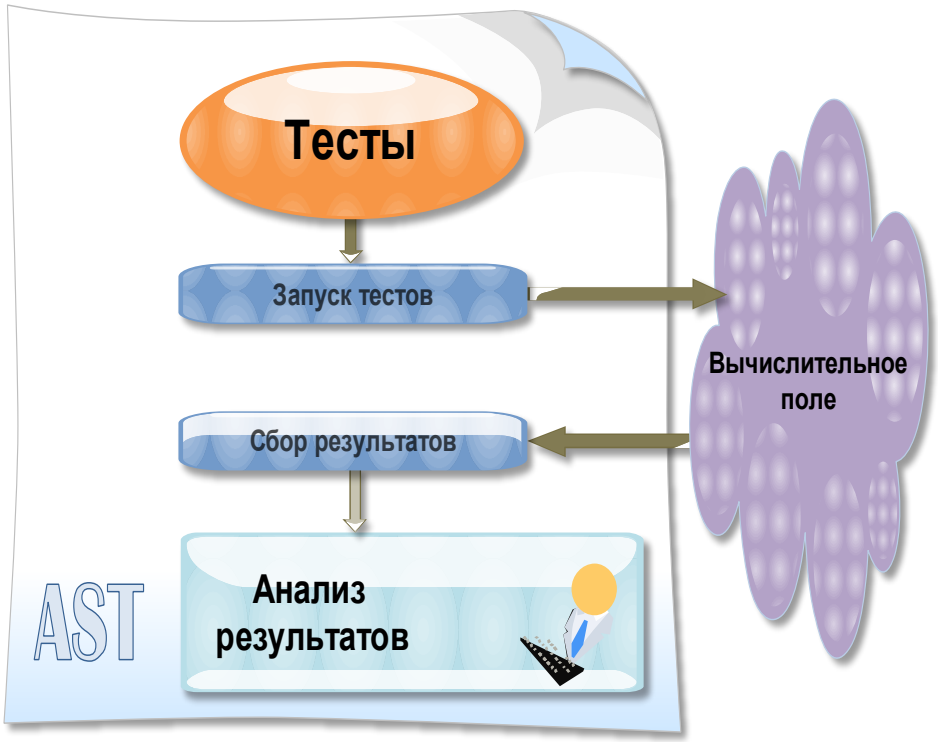
Тестовая программа ЭГИДА-ТЕСТ

Год создания:	**2012**
Назначение:	**Решение трёхмерных задач газовой динамики и теплопроводности.**
Схемы, методы:	**явные (газодинамика) и неявные (теплопроводность) разностные схемы, метод прогонок.**
Распараллеливание:	**3D геометрическая декомпозиция с 2-х слойным наложением. Нерегулярное поточечное распараллеливание. Совмещение вычислений с межпроцессорными обменами. MPI. Работа с гетерогенными системами.**
Аппаратная платформа:	**Многопроцессорные вычислительные системы с многоядерными универсальными процессорами.**
Язык:	**C++.**
Особенности:	**Интенсивная работа с оперативной памятью.**
Параметры теста:	**Размер пространственной сетки. Количество процессоров.**
Развитие:	**Использование адаптивно-встраиваемых дробных сеток. Работа с процессорами типа Intel Xeon Phi. Двухуровневое MPI+OpenMP распараллеливание. Динамическая балансировка.**

Тестовая программа MoDyS

Год создания:	**2003**
Физическая постановка:	**Установление термодинамического равновесия в кластере меди.**
Численные методы:	**Уравнения движения классической динамики Гамильтона системы материальных точек, находящихся в потенциальном поле сил межчастичного взаимодействия.**
Распараллеливание:	**MPI. С использованием технологии CUDA . Гибридный вариант –совместная работа АРУ и универсальных процессоров (часть MPI процессов работает на ядрах универсальных процессорах совместно с АРУ, а остальная часть работает на оставшихся ядрах ЦП узла без АРУ)**
Аппаратная платформа:	**Универсальные процессоры, АРУ, Intel Xeon Phi.**
Язык:	**C++, Nvidia CUDA, Фортран 90.**
Особенности:	**Динамическая балансировка.**
Параметры теста:	**Автоматическое масштабирование задачи на заданное число процессоров. Возможность задания различного количества ячеек для процессоров разной производительности.**
Развитие:	**Распараллеливание на OpenMP.**

Автоматизированная Система Тестирования



Назначение: упростить работу по запуску тестов, дальнейшему сбору результатов и их размещению в специальных структурированных хранилищах данных.

Язык программирования*: Python*.

Графический интерфейс: да.

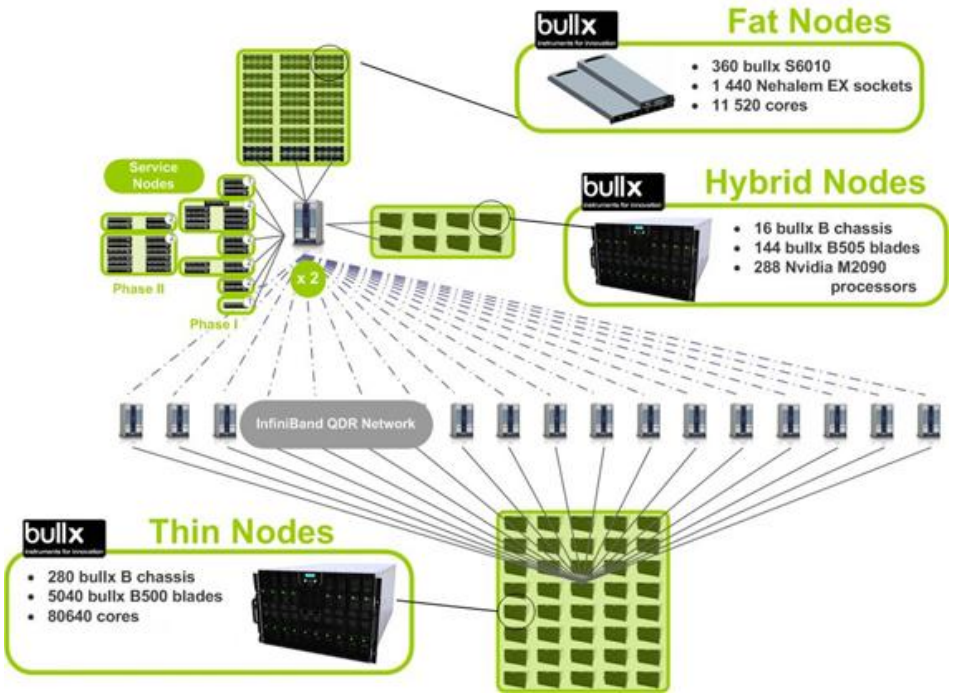
Состав пакета: тесты разработки ВНИИЭФ (tdu, pauk, gd2 и др.) и тесты сторонних разработчиков (sppm, nas, hpl и др.).

Поддерживаемые системы управления заданиями: pbs, slurm, jam.

Хранение данных: *XML*-файл или БД.

Проблемы тестирования гетерогенных и гибридных супер-ЭВМ

Архитектура гибридной гетерогенной супер-ЭВМ Curie



- Наличие узлов с универсальными процессорами различной производительности
- Наличие узлов с ускорителями или сопроцессорами
- Большое количество вычислительных устройств
- Относительно небольшое время наработки на сбой

Необходимо разрабатывать универсальные тестовые программы и соответствующую методику тестирования

HPL. Методические прикладные тесты.

Тест HPL-MAX для оценки производительности гетерогенных и гибридных супер-ЭВМ

- **HPL (High Performance Linpack), недостатки:**
  - **ориентирован на однородные системы**
  - **невозможно использовать на графических ускорителях (GPU)**
  - **требуются десятки часов бессбойной работы всей машины**
- ***HPL-MAX (Multi Architecture eXtension), разработка ВНИИЭФ:***
  - **разработан на базе HPL-2.0 (оригинальная версия), HPL-GPU-1.1.0 (Франкфуртский университет), HPL-CUDA (Nvidia)**
  - **работает на гетерогенных и гибридных системах (узлы разной производительности, включая GPU или Intel Xeon Phi), используя перераспределение нагрузки**
  - **включает реализацию механизма контрольных точек (сохранения промежуточного результата с возможностью восстановления счёта после аппаратного сбоя)**

Методические прикладные тесты

Результат: Эффективность распараллеливания, E (%)

Декомпозиция: Пространственная на параобласти с равным числом ячеек

Метод: Метод увеличения задачи (weak scaling)

Процедура тестирования:

- Расчёт на одном ядре - время T_1
- Увеличение задачи в n раз (пространственных ячеек)
- Расчёт на n ядрах - время $T_n = T_1*n+T_{накл}$
- Вычисление коэффициента ускорения - $Sp = \frac{T_{посл}}{T_n} = \frac{T_1 \cdot n}{T_n}$
- Вычисление эффективности распараллеливания - $E_n = Sp/n*100\%$

Проблема: с какого ядра (быстрого или медленного) начинать процедуру?

Два фрагмента, «быстрый» (f) и «медленный» (s)

$$T_{посл} = T_{s1} \cdot n_s + T_{f1} \cdot n_f = T_{s1} \cdot n_s + T_{s1} \cdot P_s / P_f \cdot n_f$$

$$Sp = \frac{T_{s1} \cdot \left(n_s + P_s / P_f \cdot n_f \right)}{T_n}$$

Методические прикладные тесты

Модификация Метода увеличения задачи

Статическая балансировка.

Помещаем на ядра разной производительности параобласти с различным количеством ячеек

$$T_{посл} = T_{s1} \cdot n_s + T_{f1} \cdot n_f = T_{s1} \cdot n_s + T_{s1} \cdot P_s / P_f \cdot N_f / N_s \cdot n_f$$

$$Sp = \frac{T_{s1} \cdot \left(n_s + P_s / P_f \cdot N_f / N_s \cdot n_f \right)}{T_n}$$

Произвольное количество гомогенных фрагментов

$$T_{посл} = \sum_{i=1}^{M} T_1^i \cdot n_i = T_1^b \cdot n_b + \sum_{i=1, i \neq b}^{M} T_1^i \cdot n_i \qquad T_1^i = T_1^b \cdot P_b / P_i \cdot N_i / N_b$$

$$Sp = \frac{T_1^b \cdot \left(n_b + \sum_{i=1, i \neq b}^{M} \left(P_b / P_i \cdot N_i / N_b \cdot n_i \right) \right)}{T_n}$$

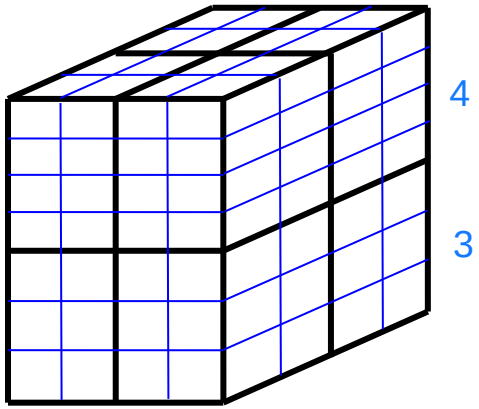
Процедура тестирования начинается с гомогенного фрагмента «b»

E_n = Sp/n*100%

Инвариантность результата относительно выбора начального фрагмента

Методические прикладные тесты

Статическая балансировка



$P_b / P_i = 3/4 = N_i / N_b$

- Сохранение топологии межпроцессорных обменов
- Сохранение регулярности пространственной сетки

Процедура тестирования

- Запуск тестов на одном вычислительном узле из каждого гомогенного фрагмента вычислительного поля. Определение $\mathbf{P_i}$
- Выбор начального гомогенного фрагмента **b**
- Определение коэффициента балансировки (отношение числа ячеек в параобластях) $N_i/N_b \approx P_b/P_i$
- Запуск теста на начальном (b) фрагменте. Затем задействование остальных фрагментов. Результат – $\mathbf{T_i}$
- Вычисление Sp и E

$$Sp = \frac{T_1^b \cdot \left(n_b + \sum_{i=1, i\neq b}^{M} \left(P_b/P_i \cdot N_i/N_b \cdot n_i \right) \right)}{T_n}$$

E_n = Sp/n*100%

Гетерогенная супер-ЭВМ с АрУ

POCATOM

Проблемы

- Сложная гетерогенная структура гибридных супер-ЭВМ.
- Большая разница в производительности узлов с АрУ и узлов с универсальными ядрами
- Отсутствие устоявшихся критериев оценки эффективности АрУ в составе узлов супер-ЭВМ и гибридной вычислительной системы в целом.
- Большая разница в архитектуре АрУ и универсальных ядер (SIMD и MIMD дисциплины), численных алгоритмах решения задачи, а также программной реализации.
- Совместное, как правило, использование прикладными программами в ходе счета универсального ядра (функции управления и вычислителя) и АрУ (функции основного вычислителя)

Решение

- Создание универсальных тестов (ЦПУ+АрУ+MIC).
  Разработаны во ВНИИЭФ: HPL, Методические прикладные тесты
- Применение модифицированного Метода увеличения задачи.
- Квант вычислительного устройства – узел Супер-ЭВМ (а не ядро).

Гибридные Супер-ЭВМ. Эксафлопс.

POCATOM

Традиционные параметры

- Общая производительность (HPL, другие тесты). 70-80%
- Эффективность распараллеливания (Методические прикладные тесты) 40-60%

Новые параметры

- Производительность к энергопотреблению.
- Производительность к цене.
- Производительность к размерам.
- Время сохранения и восстановления контрольной точки
- Время наработки на сбой
- ???

Проблемы

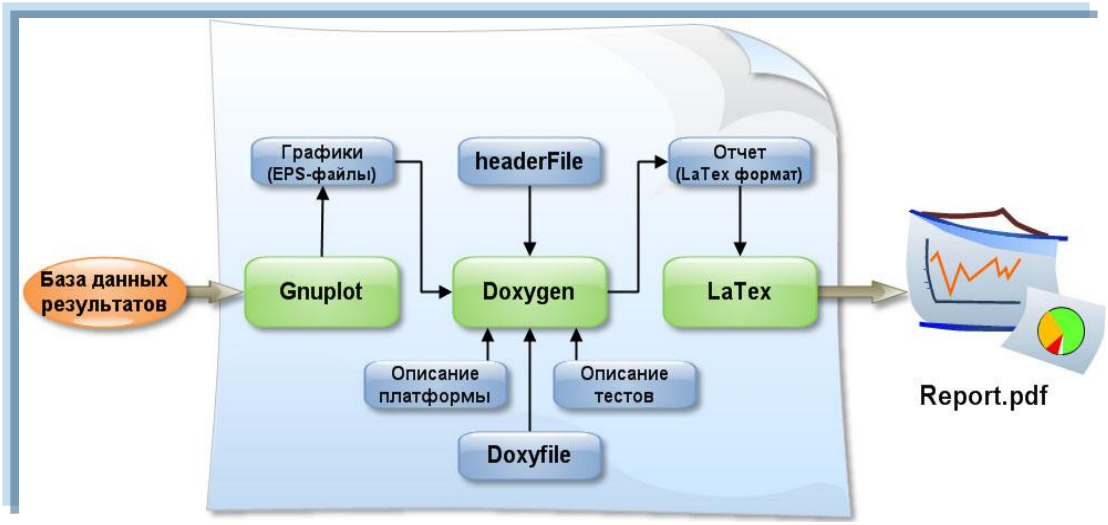
- Определение критериев.
- Новая методология тестирования.

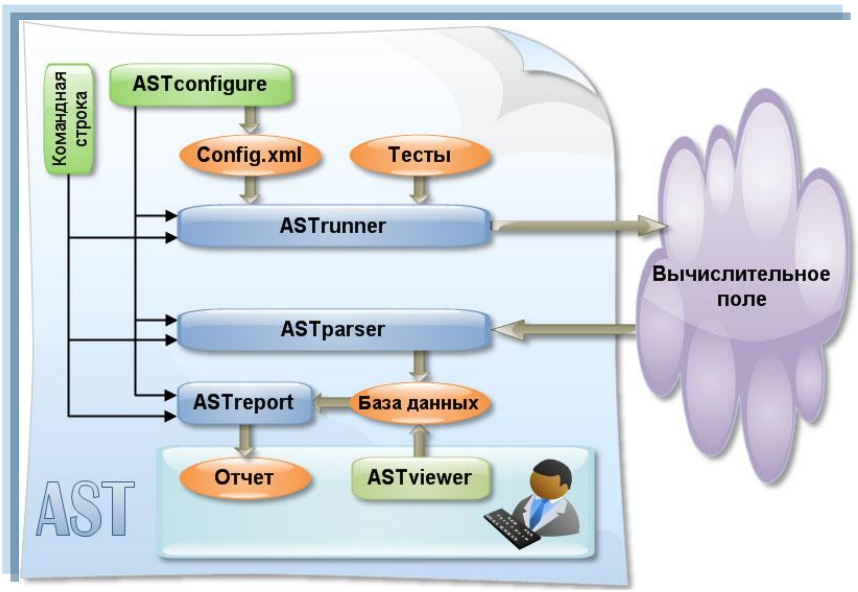
Спасибо за внимание

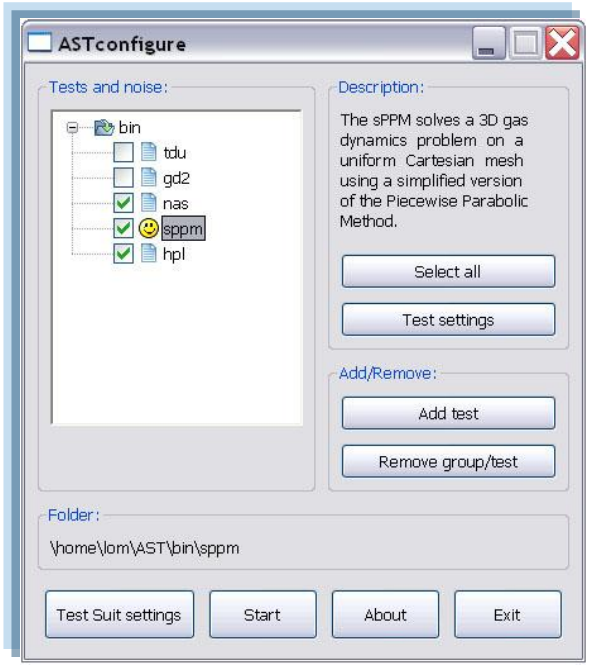
ВОПРОСЫ?

Дополнительные слайды

ACT


Графики
(EPS-файлы)
headerFile
Отчет
(LaTex формат)
База данных
результатов
Gnuplot
Doxygen
LaTex
Описание
платформы
Описание
тестов
Doxyfile
Report.pdf


Командная
строка
ASTconfigure
Config.xml
Тесты
ASTrunner
ASTparser
ASTreport
База данных
Отчет
ASTviewer
AST
Вычислительное
поле


ASTconfigure
Tests and noise:
bin
tdu
gd2
nas
sppm
hpl
Description:
The sPPM solves a 3D gas dynamics problem on a uniform Cartesian mesh using a simplified version of the Piecewise Parabolic Method.
Select all
Test settings
Add/Remove:
Add test
Remove group/test
Folder:
\home\lom\AST\bin\sppm
Test Suit settings
Start
About
Exit

Тестовая программа С-МК

	СМК	СМК-У
Год создания	1995 (С-95)	2011
Физическая постановка:	Перенос нейтронов и гамма-квантов в трёхмерном приближении	
Численный метод	Метод Монте-Карло	
Распараллеливание	MPI, асинхронное по траекториям частиц	Векторное распараллеливание
Аппаратная платформа	Универсальные процессоры	Универсальные процессоры GPU (NVIDIA)
Язык	Фортран-90	C++ и CUDA
Особенности	Нет программных ограничений на число процессоров. Использование на гетерогенном поле одновременно разных универсальные процессоры	На гетерогенном поле могут использоваться одновременно универсальные процессоры и арифметические ускорители
Параметры теста	Заказные результаты – работа с памятью. Время записи результатов – нагрузка на файловую систему Время межпроцессорного обмена – нагрузка на коммуникации	
Развитие	Арифметические ускорители (CUDA), OpenMP.	