РОССИЙСКИЙ ФЕДЕРАЛЬНЫЙ ЯДЕРНЫЙ ЦЕНТР
ВСЕРОССИЙСКИЙ НАУЧНО-ИССЛЕДОВАТЕЛЬСКИЙ
ИНСТИТУТ ЭКСПЕРИМЕНТАЛЬНОЙ ФИЗИКИ

STATISTICS TOOL KIT

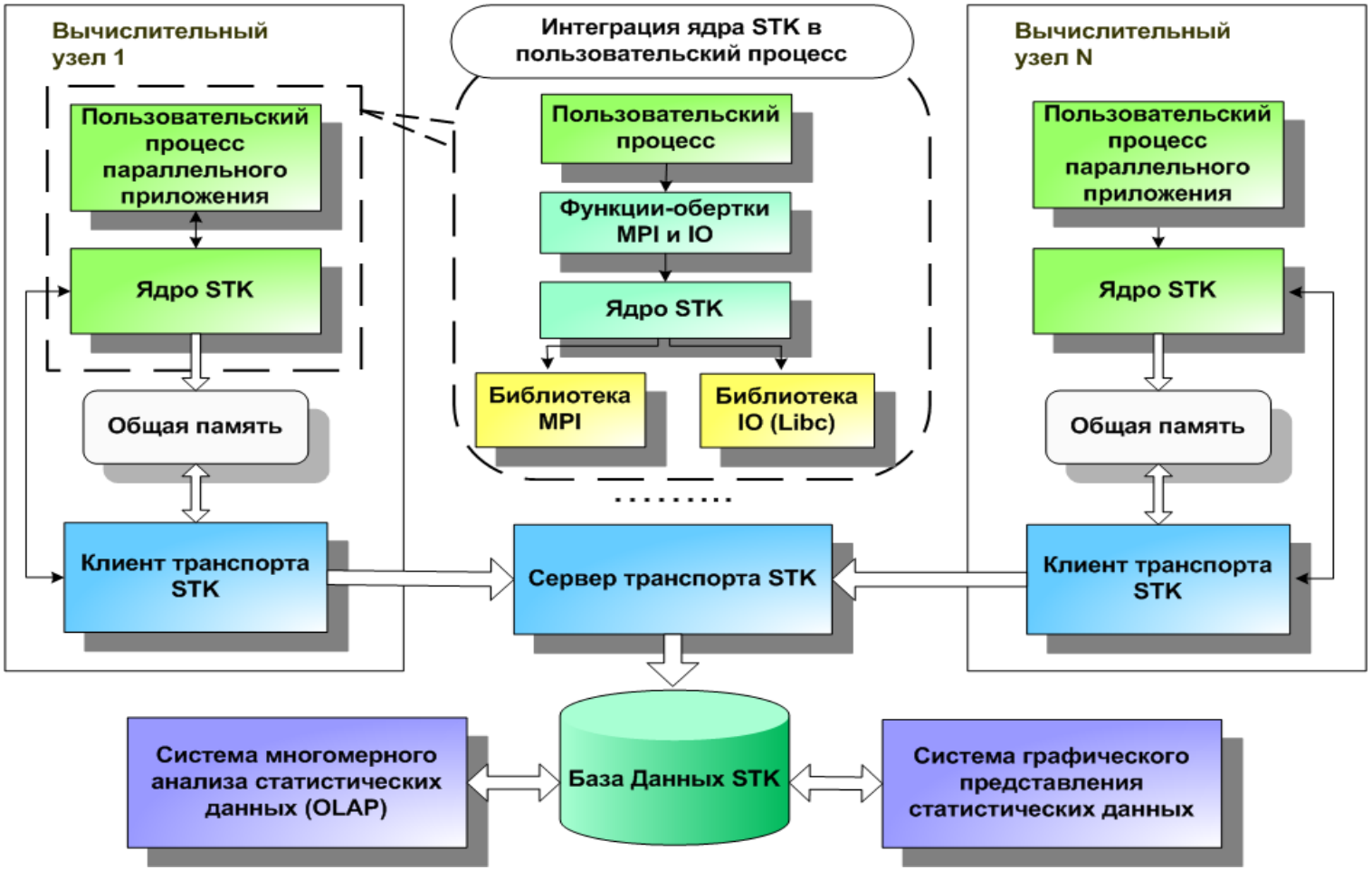
Инструментальные средства исследования эффективности параллельных приложений

Докладчик:
Новаев Д.А.
Содокладчики:
Колпаков С.И.,
Липов Д.И.

Назначение

- Сбор и представление данных по эффективности выполнения параллельного производственного счета на компонентах НВК ИТМФ ВНИИЭФ.
- Организация контроля за эффективностью выполнения параллельных задач пользователей и обеспечение более эффективного счета задач.
- Предоставление разработчикам параллельных программных комплексов средств для определения признаков и причин неэффективной работы как всей программы в целом, так и ее отдельных фрагментов.

Архитектура STK



Особенности STK

Проблемно-ориентированные библиотеки

ИТМФ РФЯЦ-ВНИИЭФ

Сторонние

Название	Назначение	Поддержка в STK
УРС-ОФ	**Единая унифицированная система расчета теплофизических свойств веществ**	**Есть**
PMLP/Parsol	**Параллельные решатели**	**Возможна**
MARX	**Расчет групповых анизотропных макроскопических констант среды**	**Возможна**
EFR	**Универсальное представление расчетных данных**	**Возможна**
… и другие, написанные на C/C++ или Fortran 77/90 для ОС Unix		
MPI (Mvapich, OpenMPI, HPMPI и др.)	**Передача сообщений**	**Есть**
UNIX Libc	**Ввод-вывод**	**Есть**
CUDA	**Расчеты на графических ускорителях NVIDIA**	**Есть**
… и другие, написанные на C/C++ или Fortran 77/90 для ОС Unix		

Группы перехватываемых вызовов

MPI	Операции для инициализации и завершения среды MPI
	Блокирующие операции передачи
	Блокирующие операции приема
	Неблокирующие операции передачи
	Неблокирующие операции приема
	Операции приема-передачи
	Операция барьерной синхронизации
	Операции для синхронизации процессов при обменах
	Коллективные операции обменов
	Возобновляемые (persistent) запросы на передачу
	Возобновляемые (persistent) запросы на прием
IO	Операции открытия/закрытия файлов
	Операции чтения из файлов
	Операции записи в файлы
	Операции со стандартным выводом
	Операции для работы с метаданными файлов (время создания, размер, права доступа, владелец и т.д.)
УРС-ОФ	Подпрограммы инициализации
	Подпрограммы расчетов

Характеристики выполнения групп вызовов:

1. **Utime** - время ЦП выполнения группы вызовов в режиме пользователя
2. **Stime** - время ЦП выполнения группы вызовов в режиме ядра
3. **Count_Call** - количество вызовов в группе, перехваченных ядром STK
4. **Walltime** - календарное время выполнения группы вызовов
5. **Max_Walltime** - максимальное календарное время выполнения функции в группе вызовов
6. **Sizebuf** - объем данных для операций передачи, файловых операций
7. **Max_Sizebuf** - максимальный объем данных при передаче между процессами пользовательской задачи для группы MPI вызовов

Оценка эффективности выполнения параллельных приложений

$$E = \frac{\sum_{i=1}^{N} T_{Calc_i}}{\sum_{i=1}^{N} T_{All_i}} \cdot 100\%$$ **(1)** , где

E - показатель эффективности выполнения параллельного приложения

N - количество процессов распараллеливания

T_{Calc} - время арифметических вычислений

T_{All} - общее календарное время выполнения

$$T_{Calc} = T_{All} - T_{MPI} - T_{IO} - T_{Idle\_Cpu}$$ **(2)** , где

T_{MPI} - календарное время на MPI

T_{IO} - календарное время на ввод-вывод

T_{Idle_Cpu} - время простоя ЦП без учета MPI и ввода-вывода

Средства визуализации данных STK

Statistics Tool Kit (Статистика по эффективности выполнения параллельных приложений)

Столбчатая диаграмма показателей средней эффективности выполнения

Интегральная статистика:

Фильтр

Название	Статус	Исп.время, ч	Исп.время ЦП, ч	Ср.эфф-ть %	MPI %	IO %	URS-OF %
Кластер1		5076		68.63	27.51	1.13	0.01
Методика1		3685		53.83	46.06	0.1	10.15
Иванов И.И.		2655					9.6
Процессов: 24 Тема: 04969.000 Заказчик: Сидоров С.С.		2290					5.8
ID_задания: 20344 Завершена: 01.09.2011		2290					5.83
Процессов: 60 Тема: 04969.000 Заказчик: Сидоров С.С.		365.27	21894.23				19.14
ID_задания: 20544 Завершена:	В счете	365.27	21894.23				19.14
Петров П.П.		1030.56	74126.38				10.7
Процессов: 72 Тема: 04969.0(		1030.56	74126.38				10.7
Методика2		893.61	32137.63				8.73
Методика3		388.69	18638.36				7.55
Методика4		15.19	697.87				3.15
Методика5		14.98	359.25				3.87
Методика6		0.2	1.29				10.19
Методика7		0.09	2.21				15.78
Методика8		17.23	1101.76				6.48
Методика9		36.81	2206.47				3.29
Методика10		24.29	388.21				10.83
Кластер2		2489.34	49720.23				1.35

Sort Ascending
Sort Descending
Columns

- Название
- Статус
- Исп.время, ч
- Исп.время ЦП, ч
- Ср.эфф-ть %
- Макс.эфф-ть %
- Мин.эфф-ть %
- MPI %
- IO %
- URS-OF %
- Время MPI, ч
- Время IO, ч
- Полезное время ЦП, ч
- Объем исп.ОЗУ, ГБ
- Объем записи на ФС, ГБ
- GPUs

- Комм. нагрузки MPI
- Статистика счета
- Графики
- Диагностика взаимоблокировки
- "Подозрительные" узлы и процессы

Иерархическое представления данных на 5 уровнях вложенности:

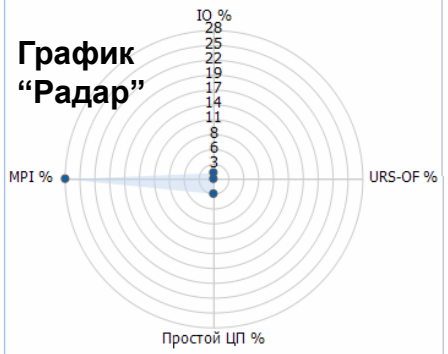
1) Многопроцессорные Вычислительные Системы;
2) Методики (Названия параллельных программных комплексов) или Группы пользователей;
3) Исполнители;
4) Количество процессов распараллеливания;
5) Задания исполнителя.

Использование библиотек:

Название: Кластер1
MPI %: 27.51
IO %: 1.13
URS-OF %: 0.01
Доля простоя ЦП %: 2.72
Набор URS-OF констант: F074F092, F076F010, U005F012, U005F095, U005F097

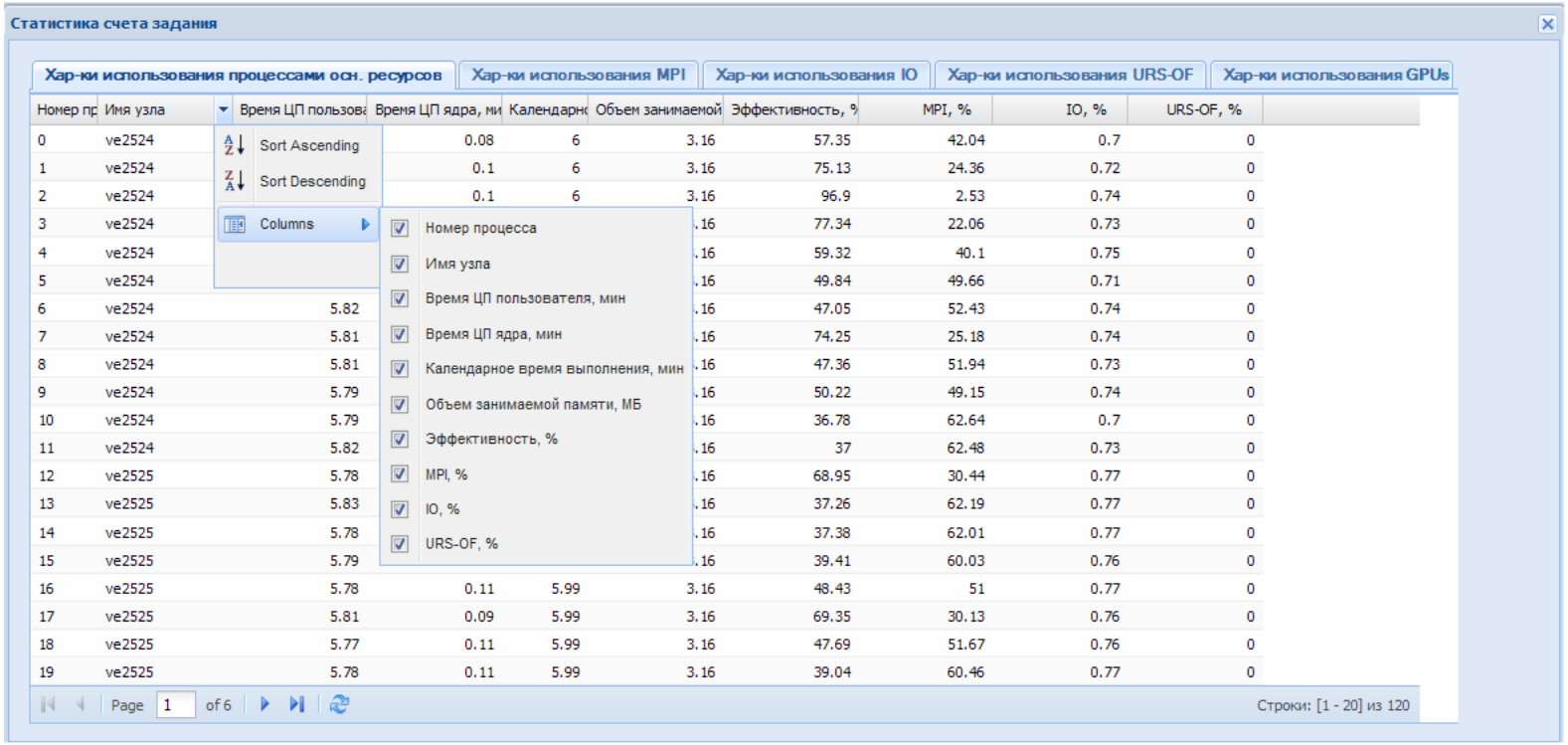
Показатели (%) простоя ЦП, использования библиотек

График "Радар"



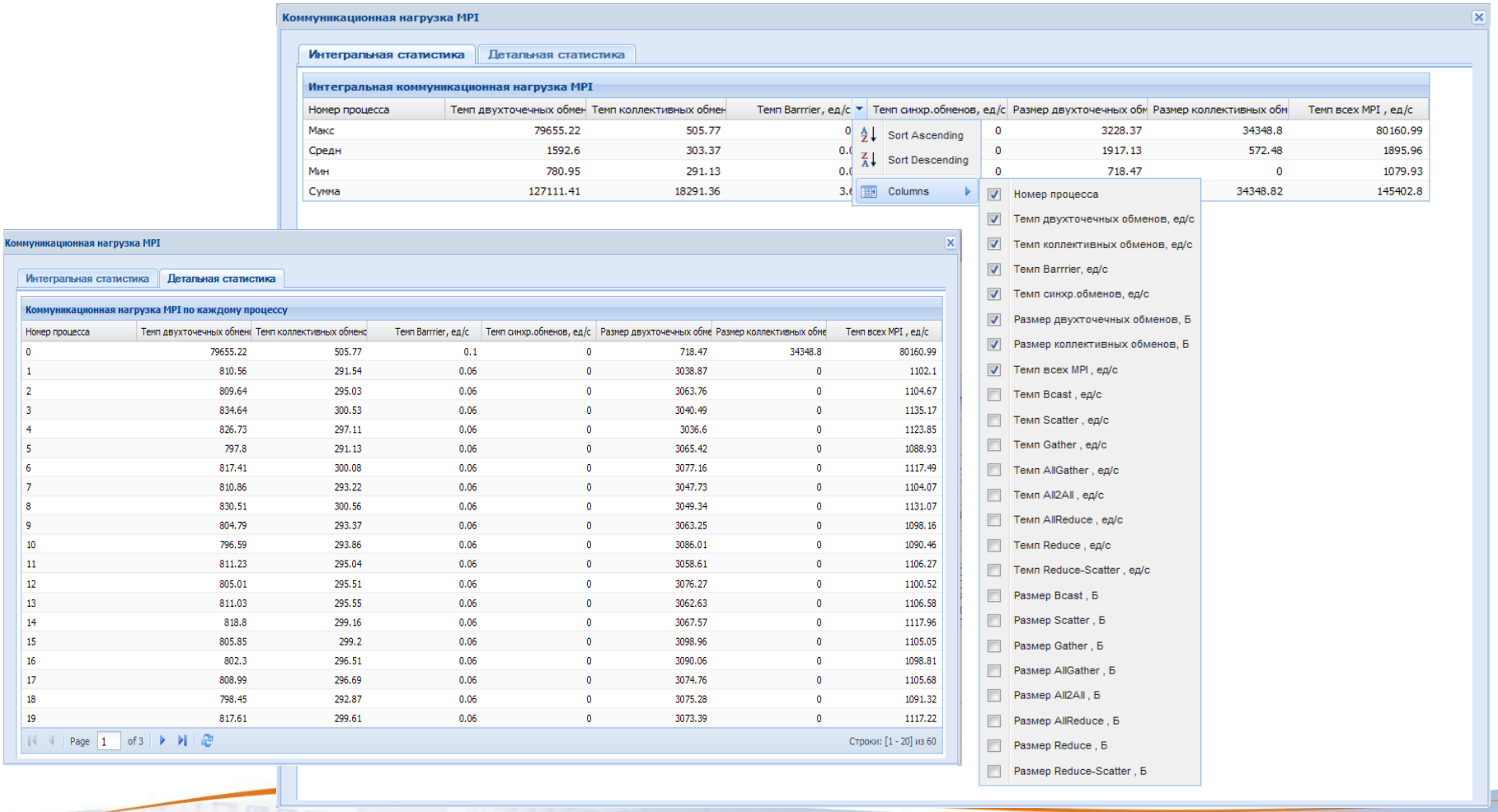
Табличное представление данных

- Характеристики использования графических ускорителей (GPUs), библиотек MPI, IO и УРС-ОФ;
- Характеристики выполнения процессов параллельного приложения (использование памяти, время выполнения, показатели эффективности, низкоуровневые метрики PAPI и др.).

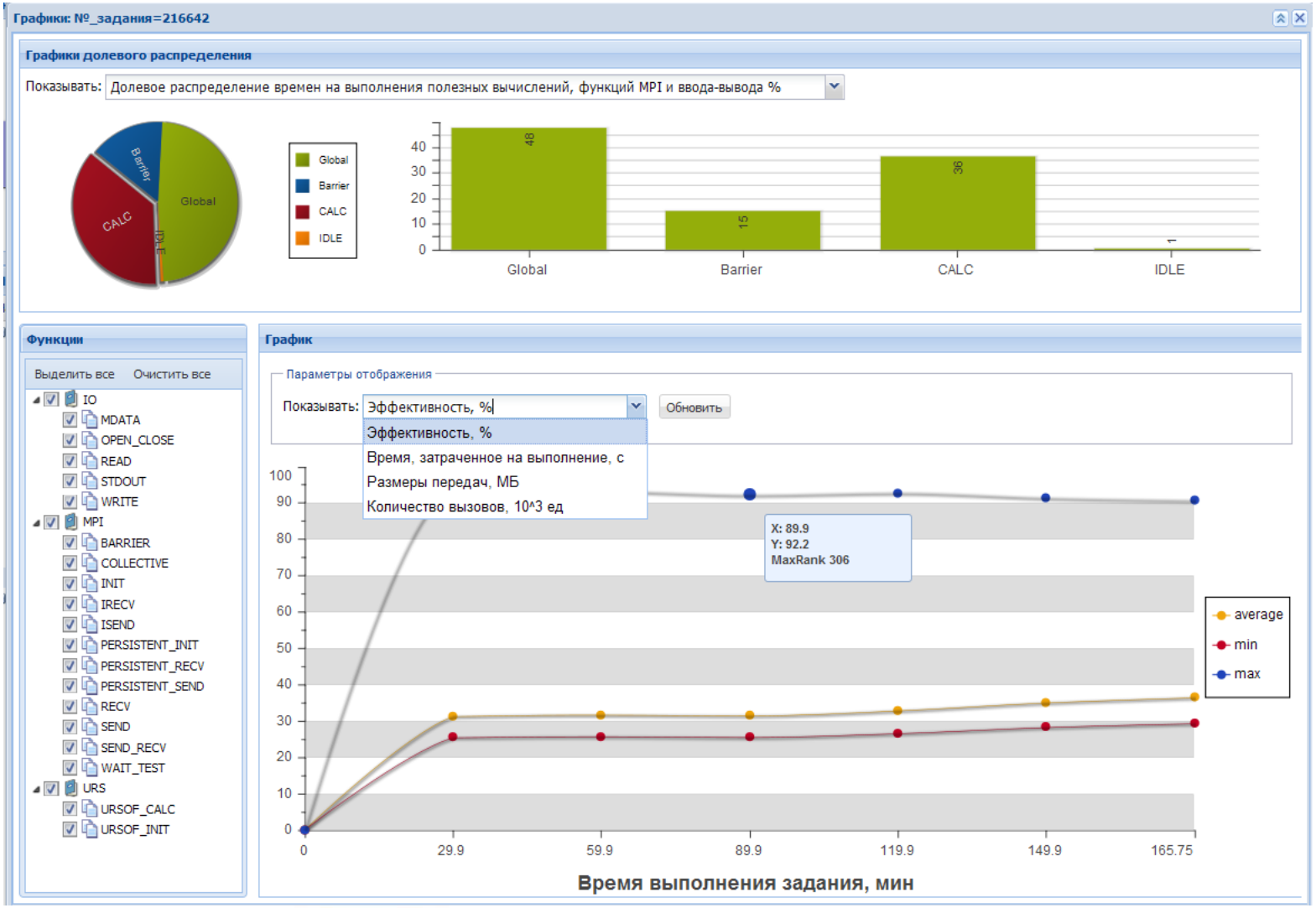


Табличное представление данных

- Коммуникационная нагрузка (темпы вызовов и объемы передач) со стороны MPI:



Графическое представление данных



Параметры использования GPU

1. ***GPU_Usage*** - процент времени относительно прошедшей секунды, в течение которого на графическом ускорителе выполнялась одна или несколько функций для GPU (kernels);
2. ***Max_GPU_Usage*** - пиковая загрузка графического ускорителя, полученная в течение времени выполнения параллельного приложения;
3. ***Mem_Usage*** - процент времени относительно прошедшей секунды, в течение которого глобальная (device) память GPU была прочитана или записана;
4. ***Max_Mem_Usage*** - пиковая загрузка при работе (чтение/запись) с памятью графического ускорителя, полученная в течение времени выполнения параллельного приложения;
5. ***Mem_Size*** - суммарное количество памяти, зарезервированное всеми активными каналами на GPU;
6. ***Max_Mem_Size*** - пиковое значение объема используемой памяти графического ускорителя, полученное в течение времени выполнения параллельного приложения.

Аппаратные счетчики современных микропроцессоров

- Набор регистров, хранящих число возникших событий:
  - *Общее число тактов;*
  - *Общее число команд;*
  - *Число операций с плавающей точкой;*
  - *Число тактов простоя функционального устройства;*
  - *Число промахов при работе с кэш-памятью;*
  - *Число промахов при работе с виртуальной памятью.*
- Отслеживание этих событий облегчает оптимизацию выполнения программы на данном CPU.

Интегральные параметры по событиям аппаратных счетчиков, представляемые STK

1. ***Мфлопс**, ед/сек* - число миллионов операций с плавающей точкой в секунду;
2. ***Флоп**, ед.* - число операций с плавающей точкой;
3. ***Обращение к кэшу L2**, ед.* - число обращений в кэш второго уровня;
4. ***Промахов кэша L3**, ед.* - число промахов кэша третьего уровня;
5. ***К эфф.памяти** (число обращений к кэшу 2-го уровня/ число промахов кэша 3-го уровня)* - коэффициент эффективности использования памяти, показывающий сколько обращений к кэшу второго уровня приходится на 1 промах кэша третьего уровня.

Заключение

- ✓ STK установлена на всех основных компонентах ВЦ РФЯЦ-ВНИИЭФ;
- ✓ STK входит в БСППО ВНИИЭФ и установлен на КС-ЭВМ разработки ВНИИЭФ, поставляемых в ряд ведущих предприятий авиа-, авто-, космической, атомной энергетики России;
- ✓ Перспективы развития:
  - сбор и представление данных по эффективности выполнения большого количества процессов (до 1000 000) параллельных задач;
  - 3D-визуализация VSTK-3D (online);
  - Интегральная статистика СУЗ (оценки работы/простоя выч.систем, параметры заданий и др.)

Спасибо за внимание !