НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Цифровой блок НИУ ВШЭ

HPC TASKMASTER – СИСТЕМА МОНИТОРИНГА ЭФФЕКТИВНОСТИ ЗАДАЧ СУПЕРКОМПЬЮТЕРА

Начальник отдела суперкомпьютерного моделирования:
Костенецкий Павел Сергеевич, к.ф.-м.н., доцент.

Собрание системных администраторов суперкомпьютерных центров России
Москва, 19.07.2021

HPC TASKMASTER – СИСТЕМА ДЛЯ ОБНАРУЖЕНИЯ НЕЭФФЕКТИВНЫХ И НЕКОРРЕКТНО ЗАПУЩЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ ЗАДАЧ

1. *Шамсутдинов А.Б., Костенецкий П.С.* Разработка системы мониторинга эффективности задач на суперкомпьютере cHARISMa // Параллельные вычислительные технологии ПаВТ'2021, 30 марта - 1 апреля 2021, г. Волгоград
2. *Костенецкий П.С., Шамсутдинов А.Б., Чулкевич Р.А., Козырев В.И.* HPC TaskMaster – система мониторинга эффективности задач суперкомпьютера // Суперкомпьютерные дни в России: труды международной конференции (27-28 сентября 2021 г., г. Москва). Москва: Издательство МГУ, 2021. В печати.

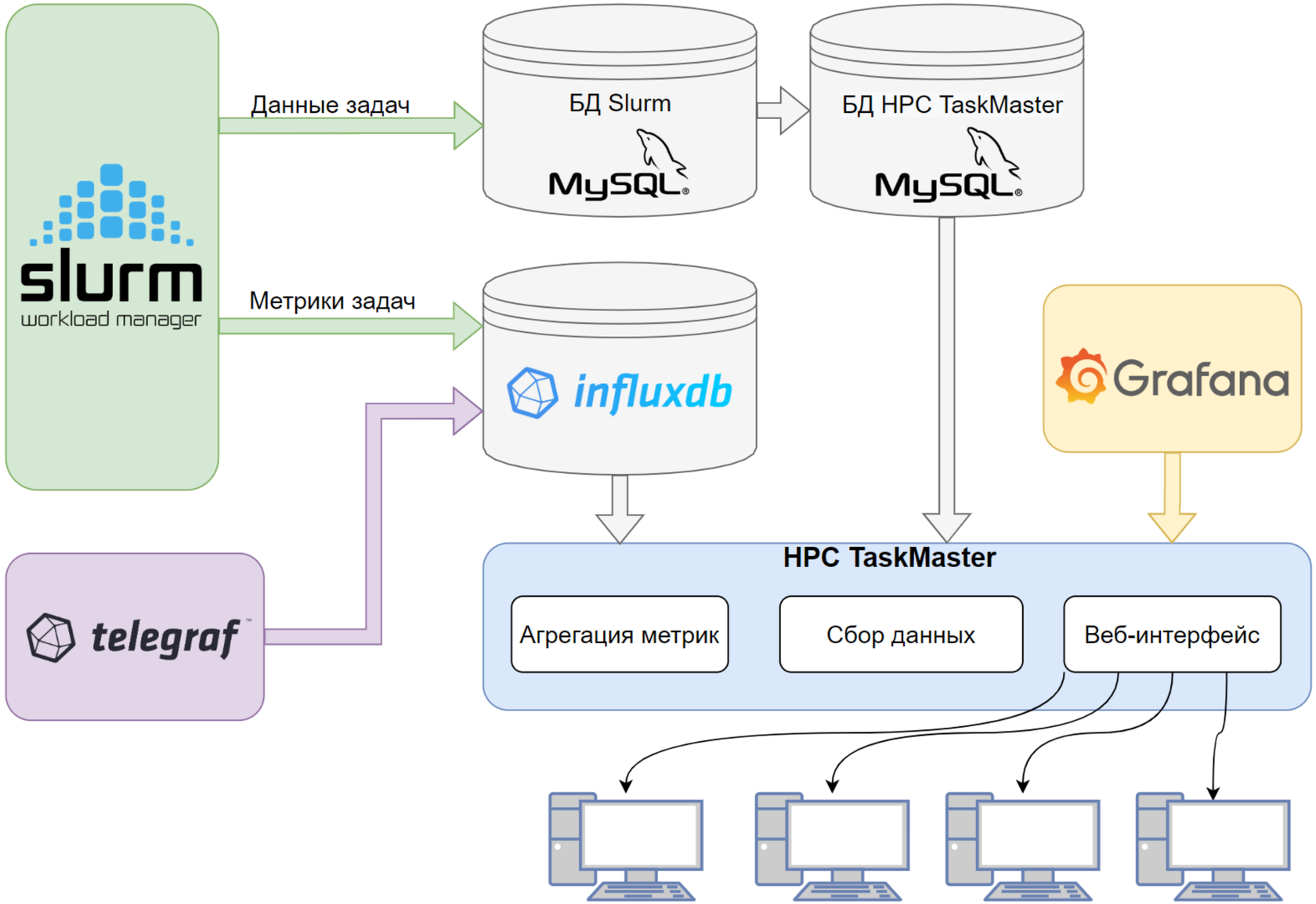
Аналоги

1. *Chan N.* A Resource Utilization Analytics Platform using Grafana and Telegraf for the Savio Supercluster. In *ACM Int. conf. proc. series*. 2019 https://doi.org/10.1145/3332186.3333053
2. *Nikitenko D. et al.* JobDigest - Detailed System Monitoring-Based Supercomputer Application Behavior Analysis // Communications in Computer and Information Science. Springer Verlag. 2017. Vol. 793. P. 516–529. DOI:10.1007/978-3-319-71255-0_42

Система позволит экономить до 30% вычислительных ресурсов.

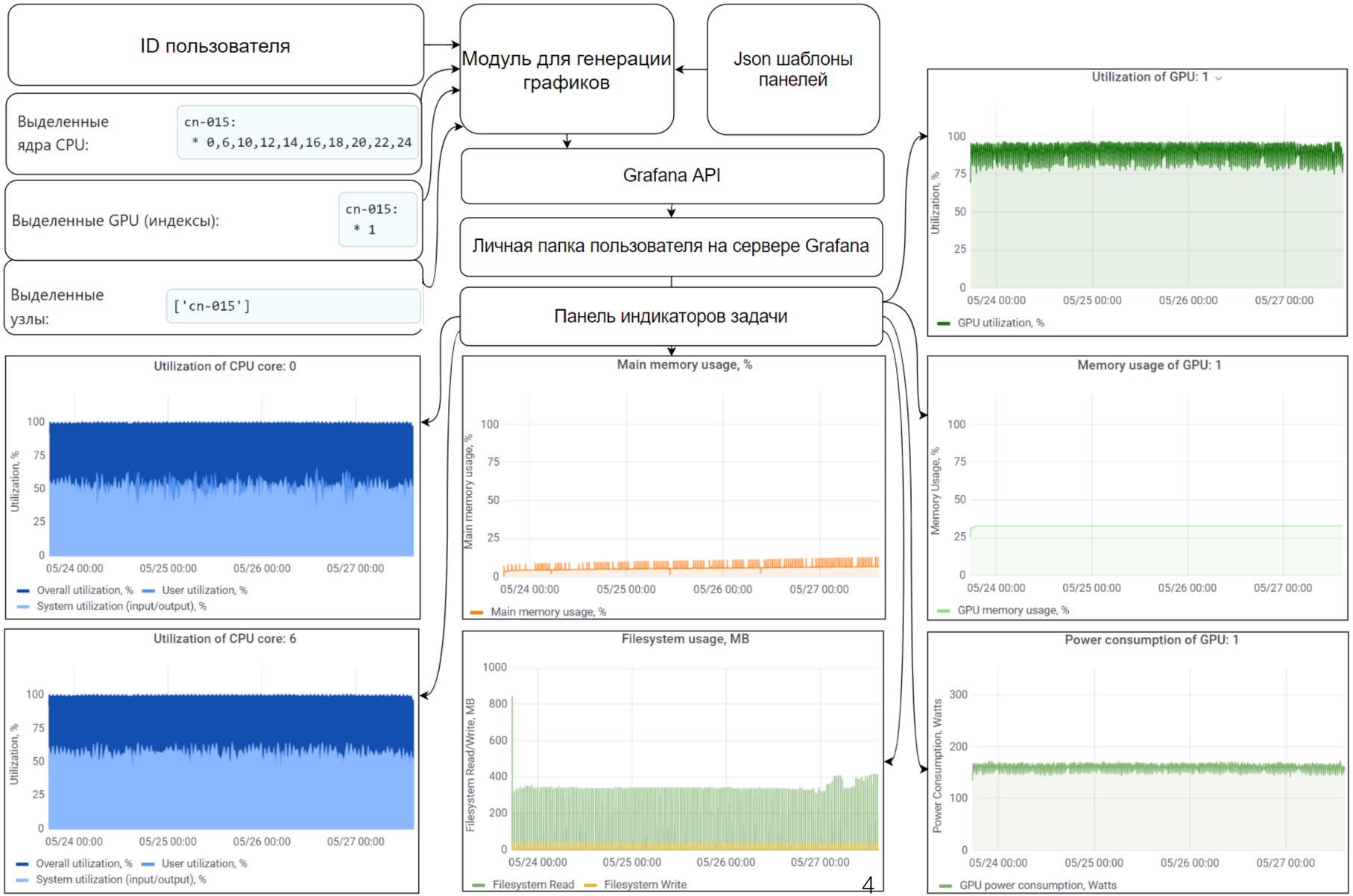
Система интегрирована в личный кабинет пользователя суперкомпьютера.

ПРИНЦИП РАБОТЫ *HPC TASKMASTER*



- Открытый исходный код
- Система собирает информацию о задачах, а не об узлах кластера
- Метрики задач автоматически анализируются на наличие проблем
- Для каждой задачи формируется вывод
- Строятся интерактивные графики при помощи Grafana

ВИЗУАЛИЗАЦИЯ ВЫПОЛНЕНИЯ ЗАДАЧИ



Для создания графиков собираются ID ядер CPU и GPU, выделенных конкретной задаче

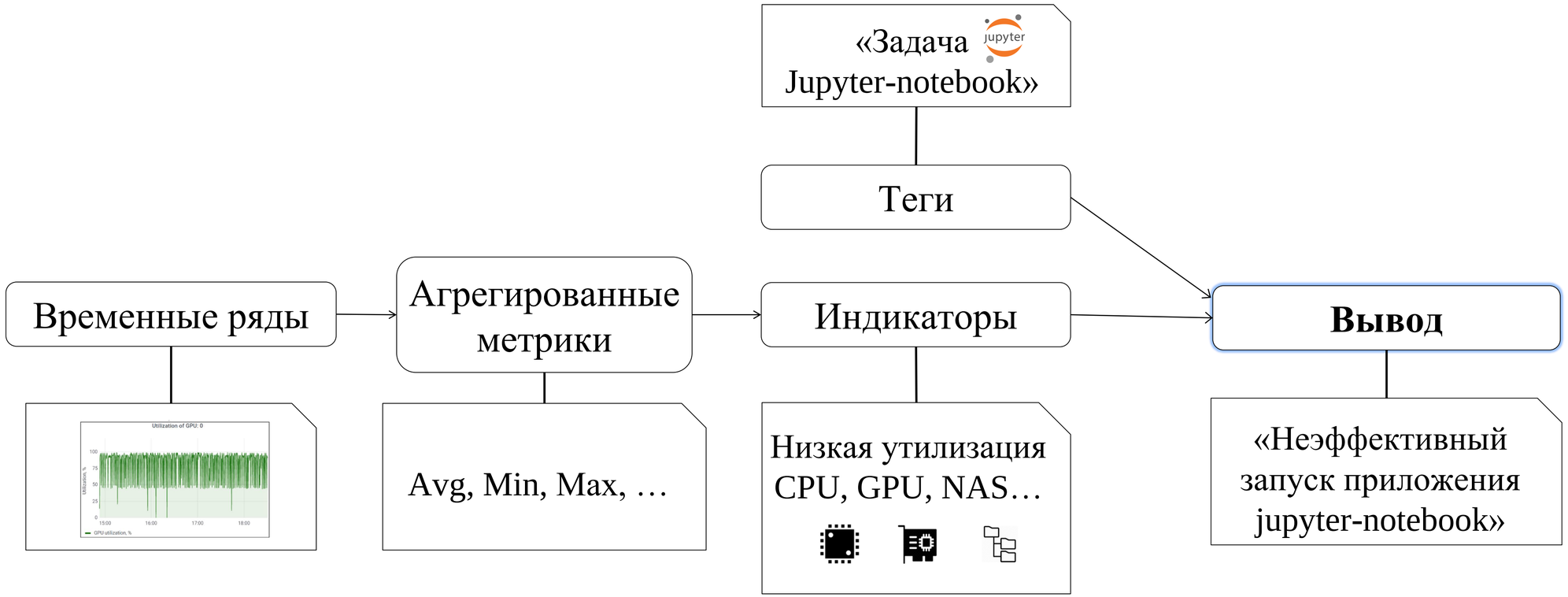
Также, собирается список вычислительных узлов, на которых запущена задача

Модуль генерирует json файл для Grafana из шаблонов и загружает его на сервер при помощи API

Для каждого пользователя создается личная папка на сервере Grafana

Отображение графиков в личном кабинете происходит при помощи технологии iframe

ЭТАПЫ ОБРАБОТКИ ДАННЫХ



- На входе есть N *временных рядов* (результаты мониторинга выполнения задач на кластере).
- По ним вычисляются *метрики* (средние значения, максимумы и минимумы и т.д.)
- Для каждой задачи получается *вектор метрик*
- Вектор метрик обрабатывается набором функций, каждая из которых выдает один *индикатор* с указанием его веса (от 0 до 1)
- Собираем все результаты функций в *вектор индикаторов* для задачи (например, «Малое использование памяти».
- Далее вектор показателей поступает на вход булевым функциям. На выходе получаются *выводы о задаче* (результата анализа, например: «Непараллельная зада запущена на нескольких ядрах»)

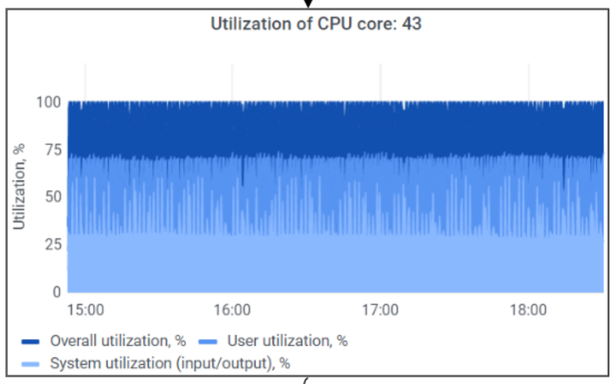
ИНДИКАТОРЫ ЗАДАЧИ

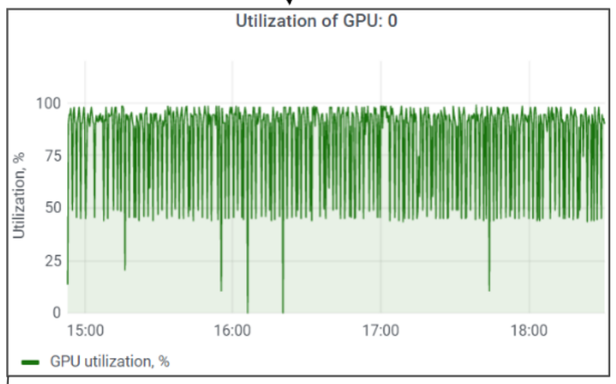
Примеры индикаторов: низкая средняя утилизация ядер CPU и GPU, низкая утилизация отдельных CPU, низкое использование видеопамяти и т.д.

Задача 298270

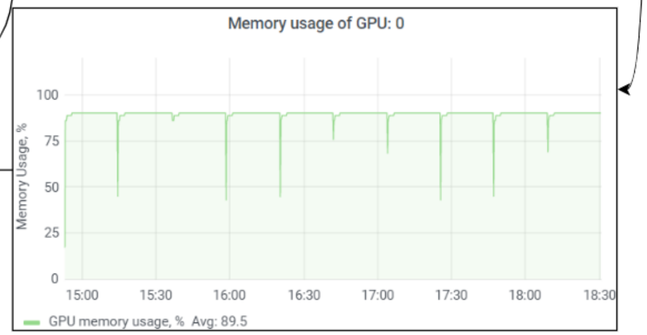
Выделенные ядра CPU:	cn-028: * 43
Задействование CPU (среднее по ядрам):	97.72 %

Выделенные GPU (индексы):	cn-028: * 0
Задействование GPU (среднее):	83.11 %
Использование памяти GPU (среднее):	28.8 GB





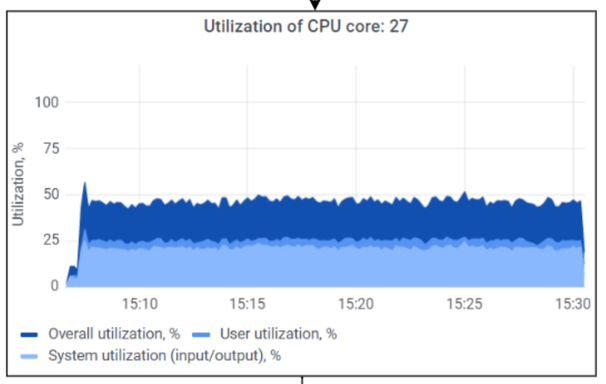
Проблем не обнаружено

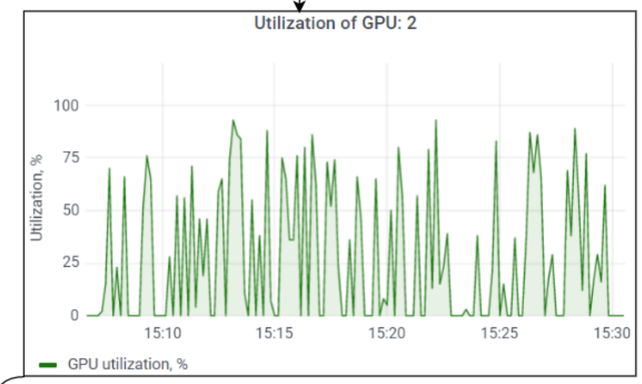


Задача 318730

Выделенные ядра CPU:	cn-010: * 27
Задействование CPU (среднее по ядрам):	44.9 %

Выделенные GPU (индексы):	cn-010: * 2
Задействование GPU (среднее):	26.95 %
Использование памяти GPU (среднее):	7.61 GB





CPU: низкая средняя загрузка:

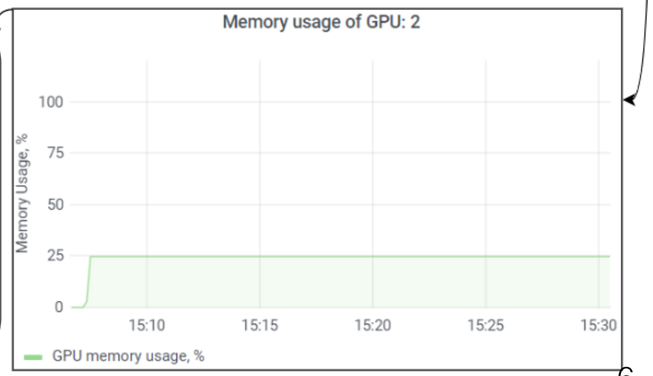
* узел cn-010: 44.9%

CPU: низкая средняя загрузка отдельных ядер:

* ядро #27(cn-010): 44.9%

GPU: низкая средняя утилизация:

* GPU IDX-2(cn-010): 26.95%



ТЕГИ ЗАДАЧИ

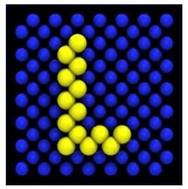
Временные теги

Задача завершилась очень быстро

Задача работает аномального долго

Теги типов задач

Jupyter Notebook



LAMMPS

VASP

GROMACS

Прочие теги

Задача завершена с ошибкой

Задача запущена на одном ядре

Выделено меньше ядер CPU, чем выделено GPU

Задача запущена на одном узле

Теги задачи – метки, передающие одно из свойств задачи

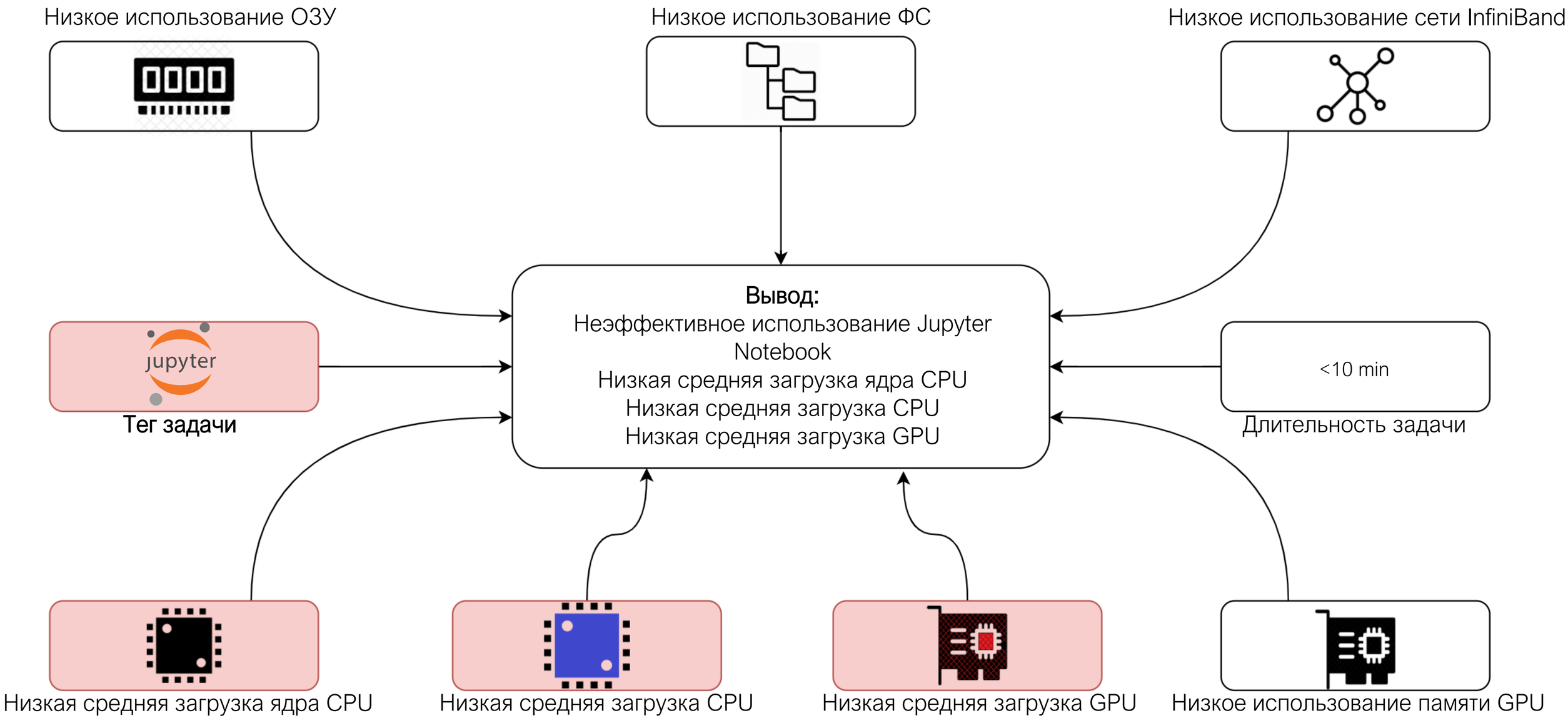
Тегами можно обозначить такие свойства, как длительность, тип задачи, обнаруженные ошибки

В отличие от индикаторов, теги не хранят в себе значение уровня

Теги нужны, чтобы делать более точные выводы о работе задачи

С помощью тегов можно с легкостью внедрять новые свойства в систему выводов

ПРИМЕР ВЫВОДА НА БАЗЕ ИНДИКАТОРОВ И ТЕГОВ

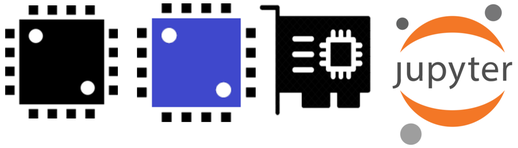


ИЕРАРХИЯ ВЫВОДОВ

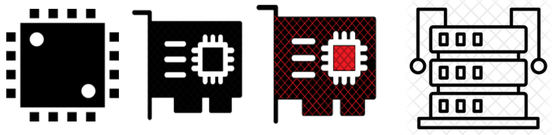
Приоритет вывода

Точные: данными выводами являются те, для которых определен сам тип задачи, что позволяет сделать точный вывод на основе индикаторов и тегов

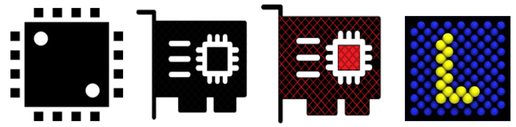
- Jupyter Notebook работает неэффективно



- Неэффективное использование srun/salloc



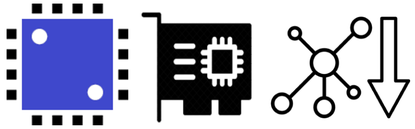
- Lammps работает неэффективно



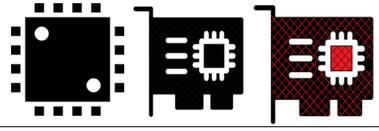
Расширенные: более глубокий вывод создается на основании нескольких индикаторов и тегов

- Ошибка в параметрах запуска задачи

- Непараллельная задача запущена на нескольких узлах

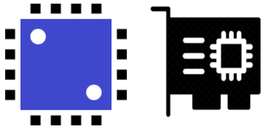


- На узле запущена непараллельная задача

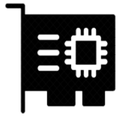


Базовые: вывод представляет собой перечисление выявленных индикаторов проблем. Такие выводы можно сделать на основании одного индикатора

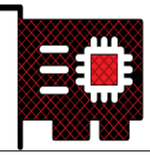
- Задача не использует CPU и GPU



- Задача не использует GPU



- Задача практически не использует память GPU



- И т. д.

ДАЛЬНЕЙШЕЕ РАЗВИТИЕ СИСТЕМЫ

- Разработка способа определения типа задач пользователей за счет анализа временных рядов при помощи математических методов и/или машинного обучения
- Добавление новых типов индикаторов и тегов для формирования новых выводов
- Разработка системы оповещений пользователей о запуске ими неэффективных задач
- Подготовка документации и предоставление открытого доступа к системе