Эффективность суперкомпьютеров: комплексный подход, основанный на исследовании исторических данных суперкомпьютера "Ломоносов-2"

Леоненков Сергей Николаевич
Жуматий Сергей Анатольевич

Факультет ВМК МГУ
НИВЦ МГУ

Актуальность проводимых исследований

Суперкомпьютер “Ломоносов-2”

Узлов:1440
Ядер: 20160
Compute

Rpeak:4,946 TFlop/s

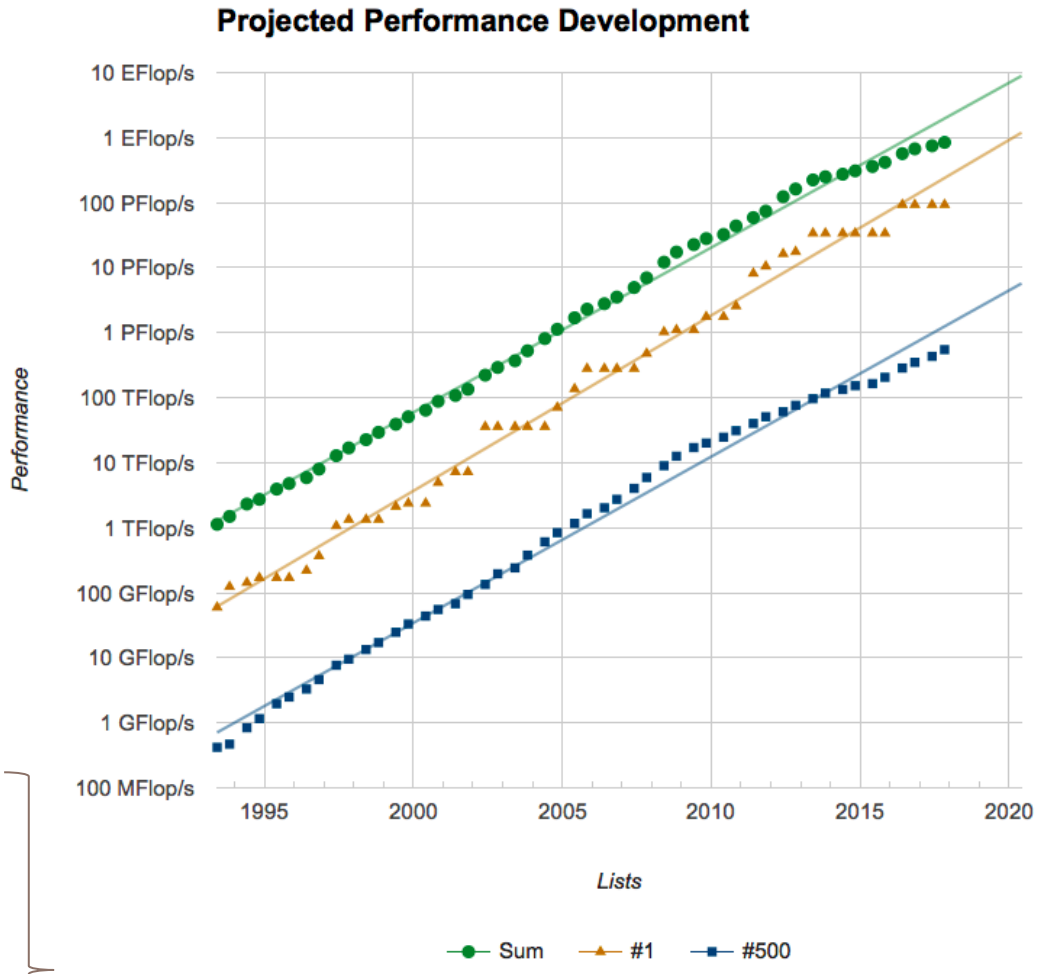
Rmax: 2,478 TFlop/s

Суперкомпьютер “Ломоносов”

Узлов:4096
Ядер: 32768
Regular4

Rpeak: 1,700 TFlop/s

Rmax: 901 TFlop/s



С каждым годом растут масштабы. (Количество ядер, пользовательских задач)

Каждый суперкомпьютерный центр определяет свою “эффективность” использования вычислительных ресурсов.

Необходим комплексный подход к организации планирования потока задач суперкомпьютерных комплексов.

Стоимость задержки работы СК

Плата за простой вычислительных ресурсов постоянно увеличивается!

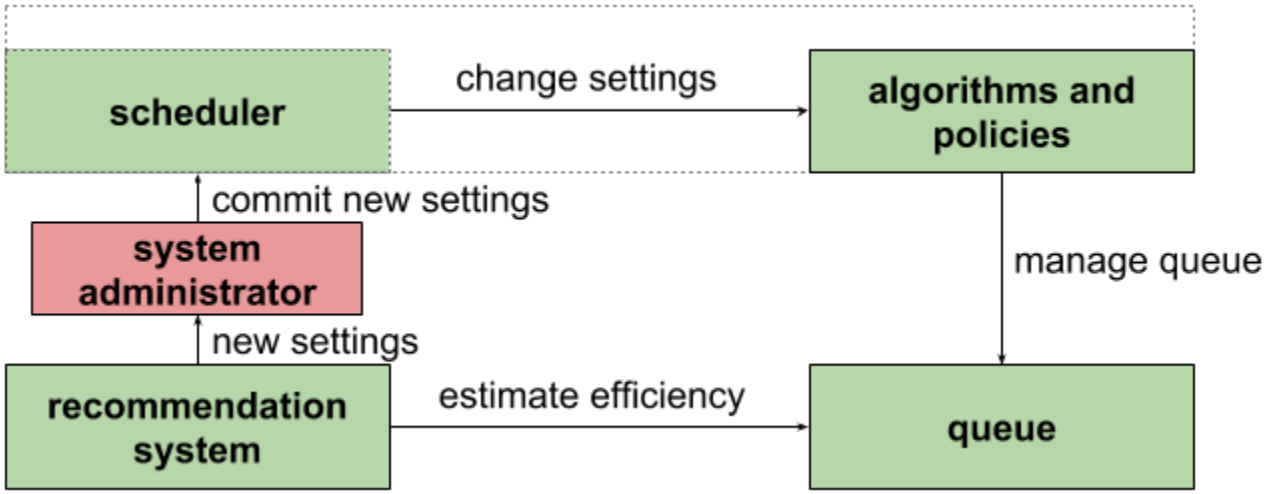
Некоторые факты:
Один день суперкомпьютера «Ломоносов» (МГУ) стоит $20 000
Один день суперкомпьютера «Titan» (ORNL, №5 в мире) стоит $300 000

Подобная ситуация везде.

Суперкомпьютер «Ломоносов»:
Если планировщик повис, половина суперкомпьютера будет простаивать уже через 2-3 часа.

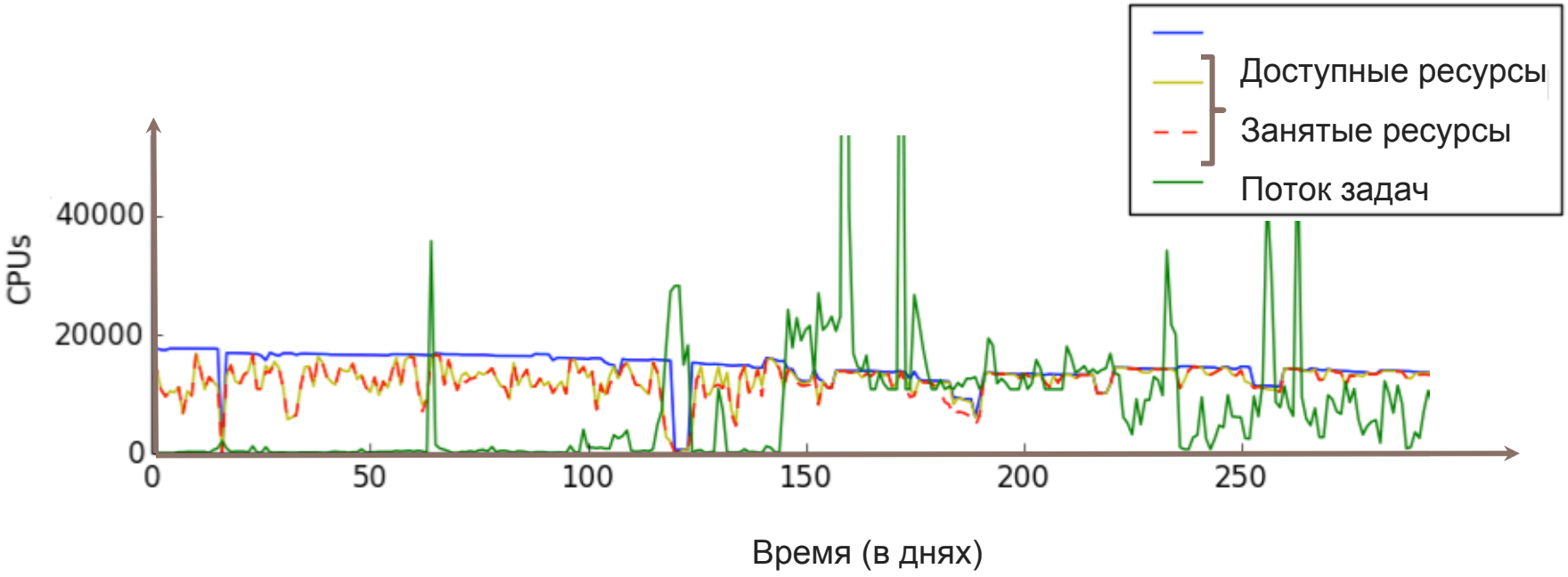
Цель исследования

Целью данной работы является исследование и разработка методов анализа эффективности планирования ***и*** использования ресурсов крупных вычислительных центров.



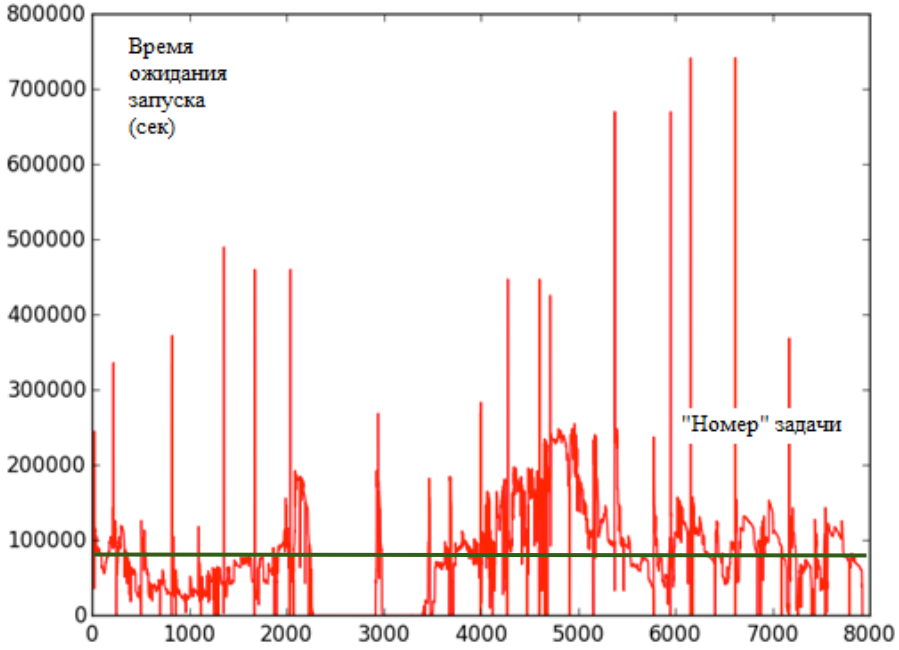
Актуальность проводимых исследований

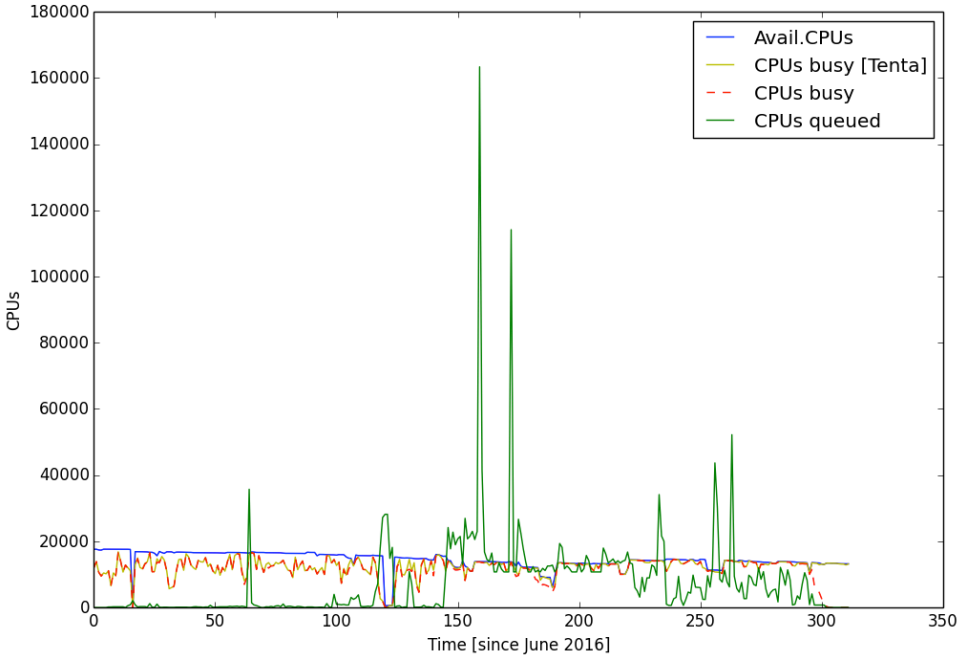
Характеристики функционирования вычислительного комплекса	
Утилизация вычислительных ресурсов	>91%
Количество обслуженных пользователей за единицу времени	35 // [15,43]
Среднее время запуска типов задач (в секундах)	small = 495.897902376 // 10059 (штук задач за период) mid = 969.612002377 // 10098 long = 1503.99341449 // 5011
Среднее время постановки задачи на исполнение	886.6 (секунд)



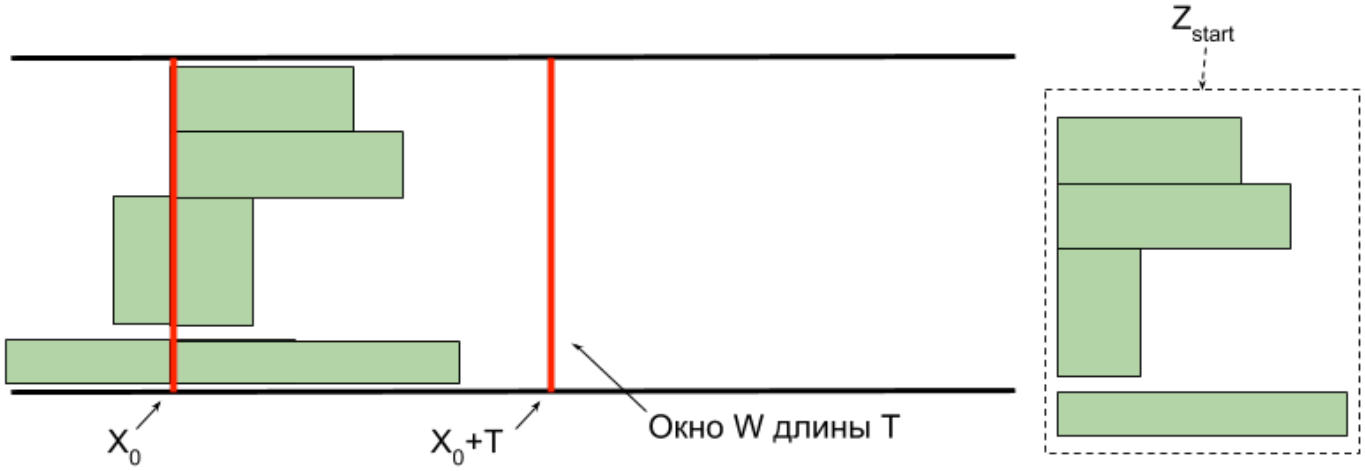
Актуальность проводимых исследований

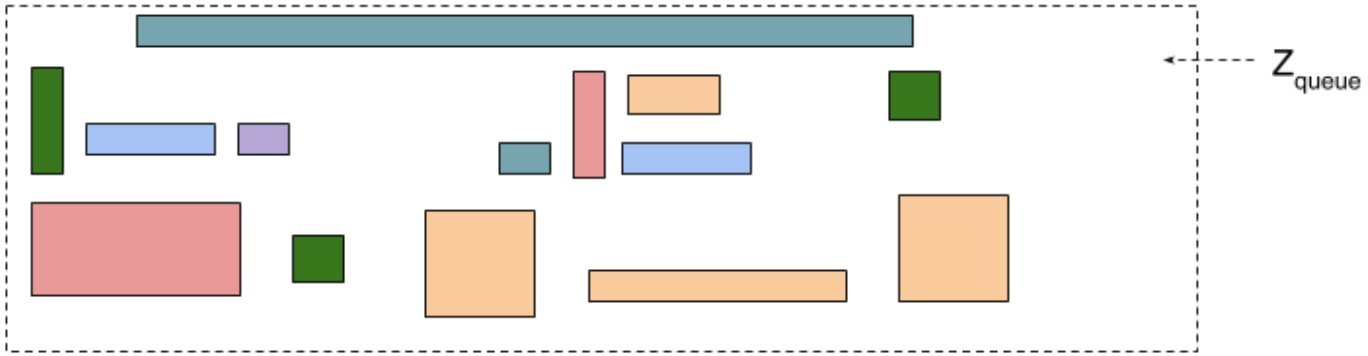
Характеристики функционирования вычислительного комплекса	
Утилизация вычислительных ресурсов	>91%
Количество обслуженных пользователей за единицу времени	35 // [15,43]
Среднее время запуска типов задач (в секундах)	small = 495.897902376 // 10059 (штук задач за период) mid = 969.612002377 // 10098 long = 1503.99341449 // 5011
Среднее время постановки задачи на исполнение	886.6 (секунд)





Вводимые понятия





Характеристики С/К

Пусть:

- UserNum(Z) — число пользователей, чьи задачи принадлежат упаковке Z.

Характеристика: **Утилизация вычислительных узлов.**

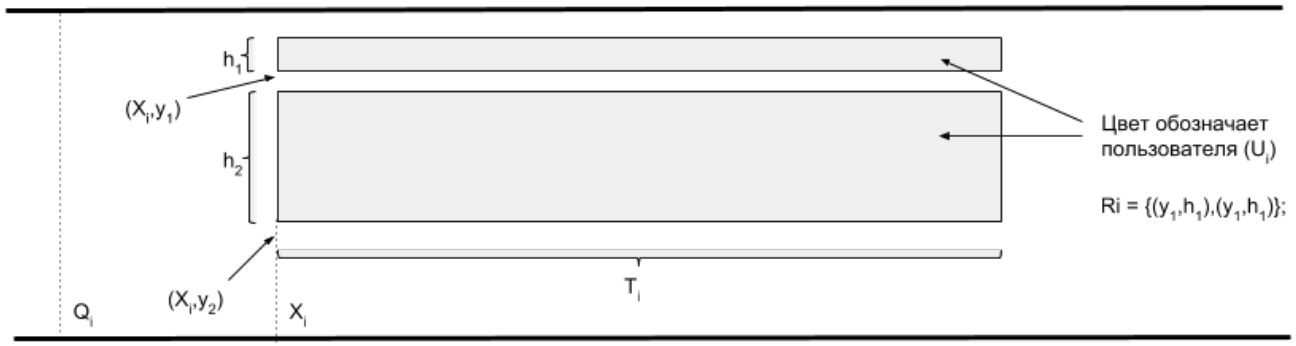
Смысл: Минимизация количество свободных ресурсов. (= уменьшаем размер “дырок в упаковке”/ нераспределенных ресурсов).

$$\text{Opt(Z,W)} = \text{Utilization(Z,W)} = 1 - \sum_{i=1}^{|Z|} (H_i * (min(T, X_i + T_i) - X_i)/(H * T);$$

Характеристика: **Среднее время старта первого задания пользователей в окне W.**

Смысл: минимизация среднего расстояния от задачи каждого цвета, у которой Xi - X0 минимально среди задач данного цвета, до начала окна W – X0.

$$\text{Opt(Z,W)} = \text{FUJStartTime(Z,W)} = \sum_{user=1}^{UserNum} min_{j \subset UserJobs(user)} (X_j - Q_j)/UserNum(Z);$$



Эффективность планирования

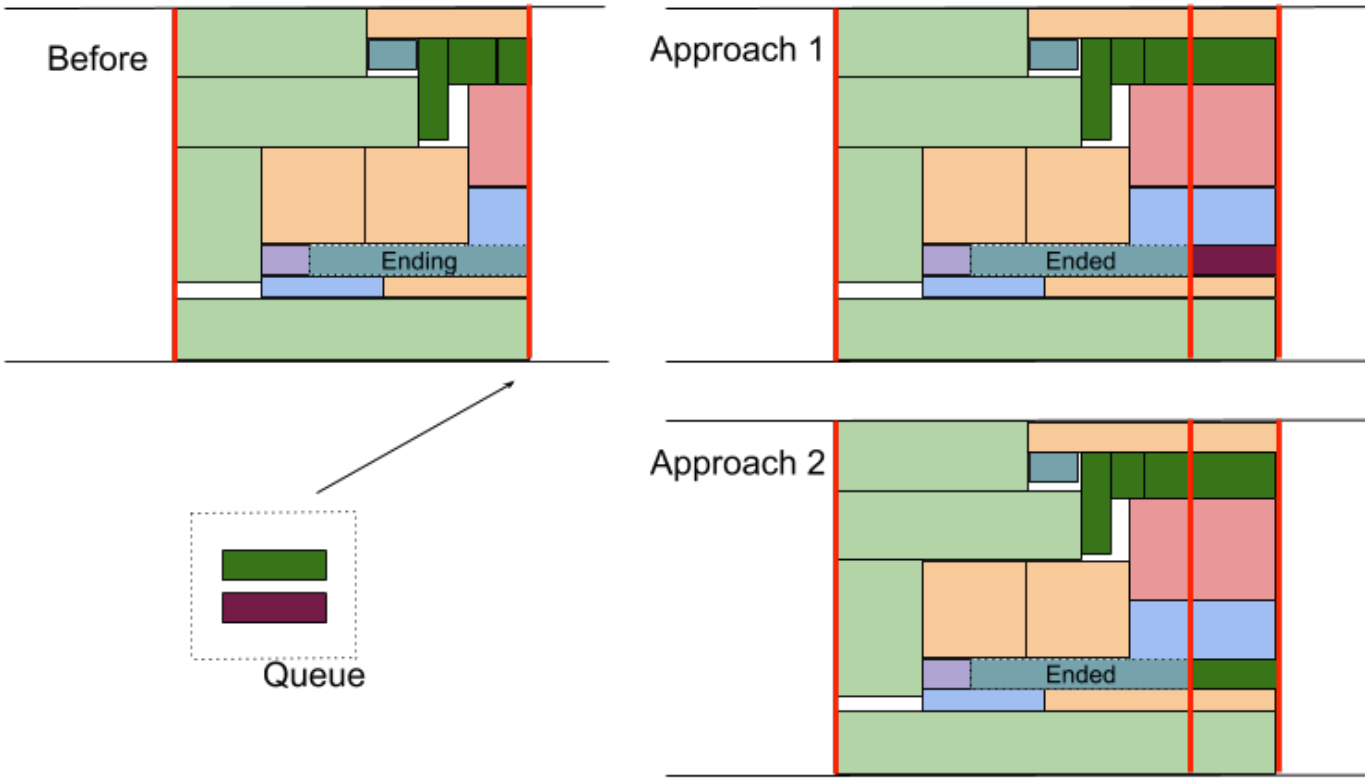
Характеристики функционирования вычислительного комплекса
Утилизация вычислительных ресурсов
Скорость запуска исполнения первой задачи с момента ее постановки в очередь для каждого пользователя
Количество запущенных задач за единицу времени
Количество обслуженных пользователей за единицу времени
Среднее время до предсказанного старта ближайшей задачи очереди
Среднее время запуска типов задач

Настройки (параметры) объектов
Алгоритм планирования (FCFS/тип Backfill) + (разные методы предсказания продолжительности задач)
Тип системы приоритетов
Лимит узло-часов за период
Лимит на количество задач в очереди (Лимит на задачи на счете + в очереди)

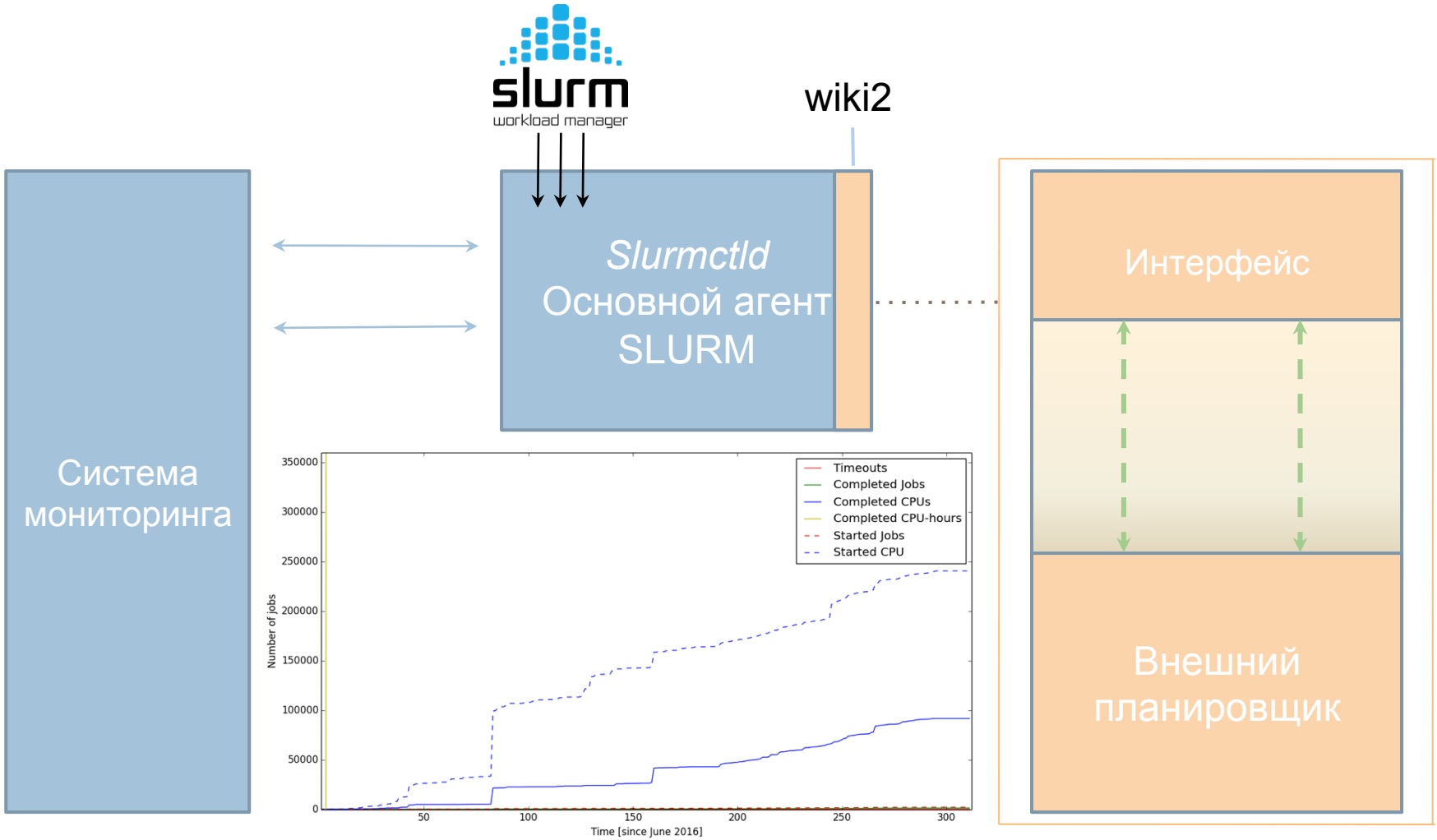
$$Efficiency = \sum_{i=1..5} (PriorityCoefficient_i * MetricsValue_i), \text{ where}$$

$$\sum_{i=1..5} (PriorityCoefficient_i) = 1.$$

Показательный пример



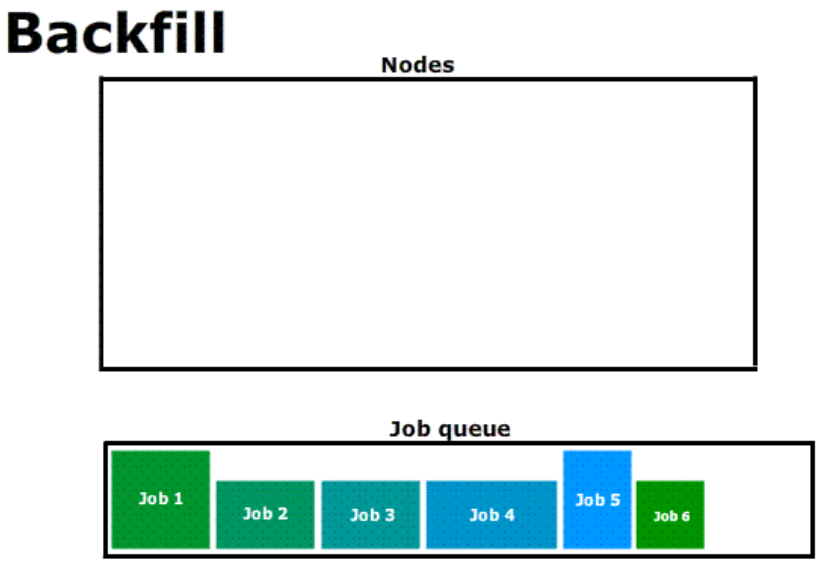
Архитектура программного комплекса



Спасибо за внимание!

Возможности внешнего планировщика

a. Внешний планировщик для SLURM ;
   i. Алгоритмы планирования;
   ii. Новые лимиты;
   iii. Приоритеты;
   iv. Поддержка резерваций;
   v. Возможности настройки личных лимитов/ приоритетов;
b. Сбор статистики по всей работе планировщика;
c. Slurm Simulator Playground;
d. Scheduling Simulator;

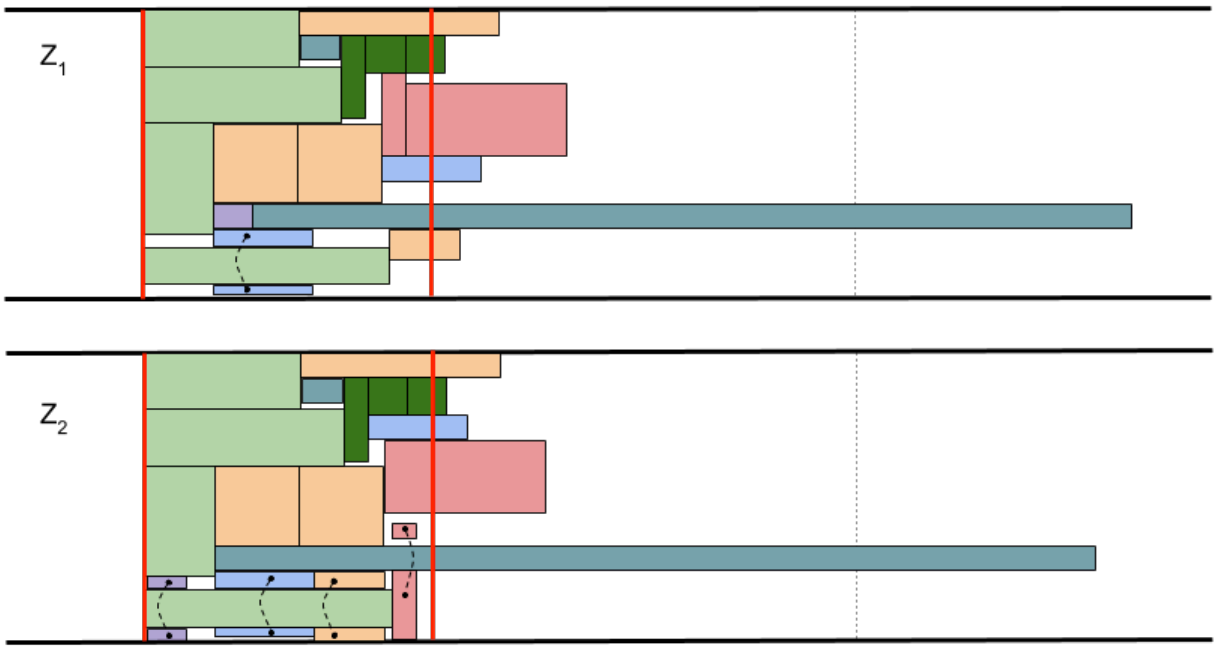


Эксперименты	Тесты на исторических данных	
	Общее использование проц. времени	Задержка старта задач
Стандартный алгоритм	0,998	1
Улучшенный алгоритм	1	0,907

Внешний планировщик

Задача упаковки

Подзадача: на заданном окне и наборах заданий Zstart и Zqueue найти упаковку заданий Zend в окно W с минимальным значением функции потери качества этой упаковки — Opt(Z,W).



Публикации и выступления

Промежуточные результаты были представлены на конференциях: International Young Scientists Conference “Computer Modeling and Simulation”, ”Методы суперкомпьютерного моделирования” в г. Таруса, ПаВТ, “Ломоносовские чтения”.

NEW: Одобрена заявка на участие в Russian Supercomputing Days 2018. Устный доклад + Статья.

Публикации:

1. Леоненков С.Н., Жуматий С. А., Алгоритмы планирования и эффективность использования суперкомпьютера «Ломоносов» // Вычислительные технологии в естественных науках: Методы суперкомпьютерного моделирования: сборник научных статей. Часть 4. М.: ИКИ РАН.
2. S. N. Leonenkov, S. A. Zhumatiy, Introducing new backfill-based scheduler for SLURM resource manager // Procedia Computer Science, Volume 66, Pages 661-669. [SCOPUS/ВАК/WoS]
3. Леоненков С.Н., Жуматий С. А., Оптимизация алгоритма Backfill и системы планирования задач для использования на суперкомпьютере “Ломоносов” // Вычислительные технологии в естественных науках: Методы суперкомпьютерного моделирования: сборник научных статей. Часть 3.
4. S. N. Leonenkov, S. A. Zhumatiy, Supercomputer Efficiency: Complex Approach Inspired by Lomonosov-2 History Evaluation // Russian Supercomputing Days 2018.

Педагогическая практика

Подготовил и прочитал лекцию по системам и работе в Суперкомпьютерном центре МГУ в рамках спецкурса кафедры СКИ под руководством Нины Николаевны Поповой.

Осуществлял поддержку работы студентов спецкурса кафедры СКИ на установках Суперкомпьютерного центра МГУ.

Принял участие в составлении и технической поддержке проведения экзамена по курсу “Суперкомпьютерное моделирование” под руководством Нины Николаевны Поповой.

Стоимость задержки работы СК

Плата за простой вычислительных ресурсов постоянно увеличивается!

Некоторые факты:
Один день суперкомпьютера «Ломоносов» (МГУ) стоит $20 000
Один день суперкомпьютера «Titan» (ORNL, №5 в мире) стоит $300 000

Подобная ситуация везде.

Суперкомпьютер «Ломоносов»:
Если планировщик повис, половина суперкомпьютера будет простаивать уже через 2-3 часа.

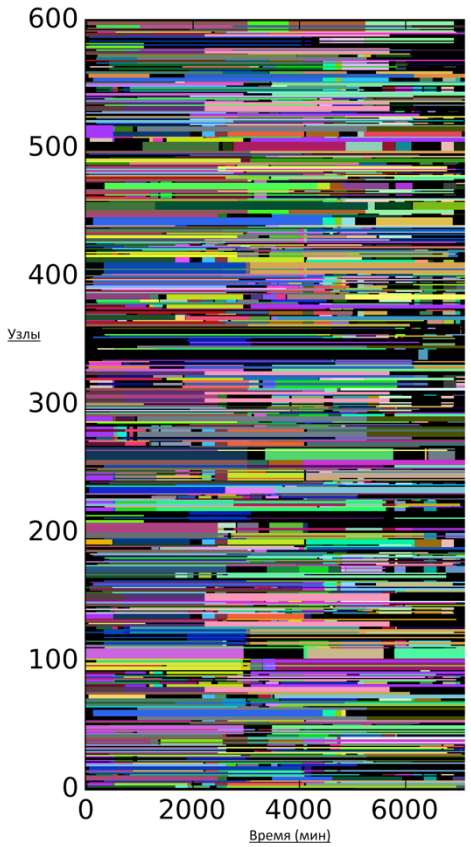
Понятия, используемые в докладе

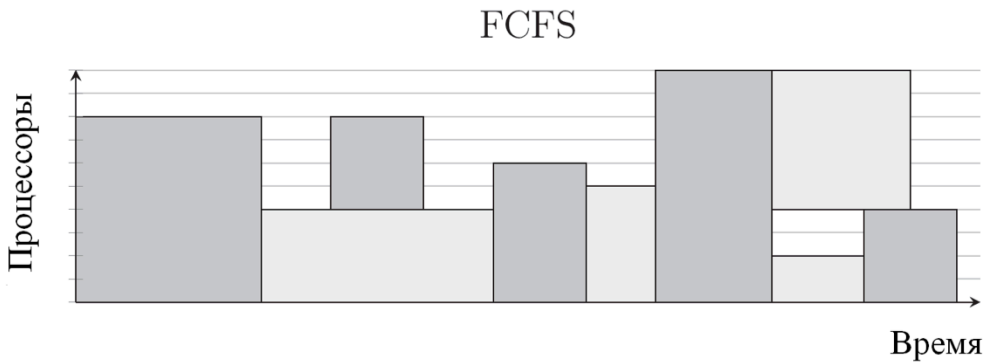
Задача — программа пользователя, поставленная в очередь на исполнение.

Поток задач — совокупность задач, которые были поставлены в очередь на исполнение.

Планирование потока задач — это определение порядка запуска задач и выбора узлов для исполнения каждой из них. Для определения порядка запуска задач используются алгоритмы планирования, например, FCFS или Backfill.

Характеристики функционирования суперкомпьютерного комплекса — совокупность свойств процесса прохождения потока задач на исполнение на конкретной суперкомпьютерной системе. Исходя из опыта использования суперкомпьютерных систем мы выделяем 6-7 основных характеристик.





Направления дальнейших работ

1. Пример машинного обучения с привязкой к “процессорному времени” и анализу эффективности его использования отдельными программами и пользователями.

2. Тесты на суперкомпьютере “Ломоносов-2” в “онлайн” режиме.

Log	EASY	EASY-Clairvoyant
KTH-SP2	92.6	71.7 (22%)
CTC-SP2	49.6	37.2 (25%)
SDSC-SP2	87.9	70.5 (19%)
SDSC-BLUE	36.5	30.6 (16%)
Curie	202.1	69.9 (65%)
Metacentrum	97.6	81.7 (16%)