

Суперкомпьютерные дни в России-2025,
Москва, 29-30 сентября 2025

Simulation model of a crowdsourcing platform for assessing the effectiveness of task distribution algorithms and results aggregation methods

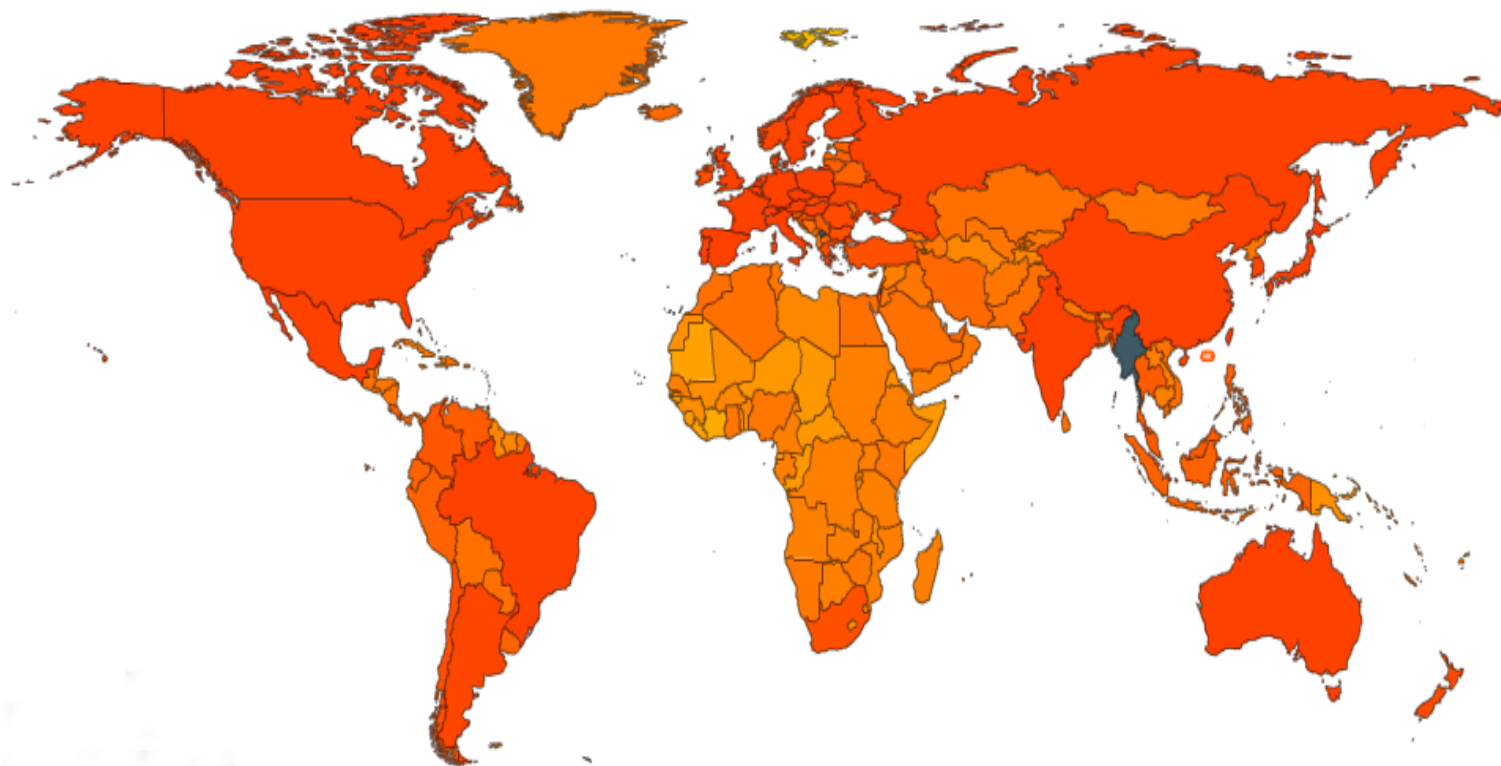
Имитационная модель краудсорсинговой платформы для оценки
эффективности алгоритмов распределения задач
и методов агрегирования результатов

А.М. Иванов¹, И.И. Курочкин¹²

¹НИТУ МИСИС, Москва,
²ИППИ РАН, Москва.

Публичные распределенные вычислительные системы

- Миллионы участников;
- Десятки миллионов устройств (вычислительных узлов);
- Больше 100 международных проектов добровольных распределенных вычислений.



Кто такие добровольцы?

Доброволец предоставляет свои вычислительные ресурсы для расчета научных распределенных проектов

Почему?

- Желание помочь науке
- Причастность к научным открытиям (желание разобраться в получаемых результатах)
- Спортивный интерес (кто больше наберет баллов)
- Общение
- Участие в социально значимых проектах

Проект Folding@home

Проект по моделированию процессов свёртывания молекул белков.

Организатор: Стэнфордский университет.

Активен с 2000 года.

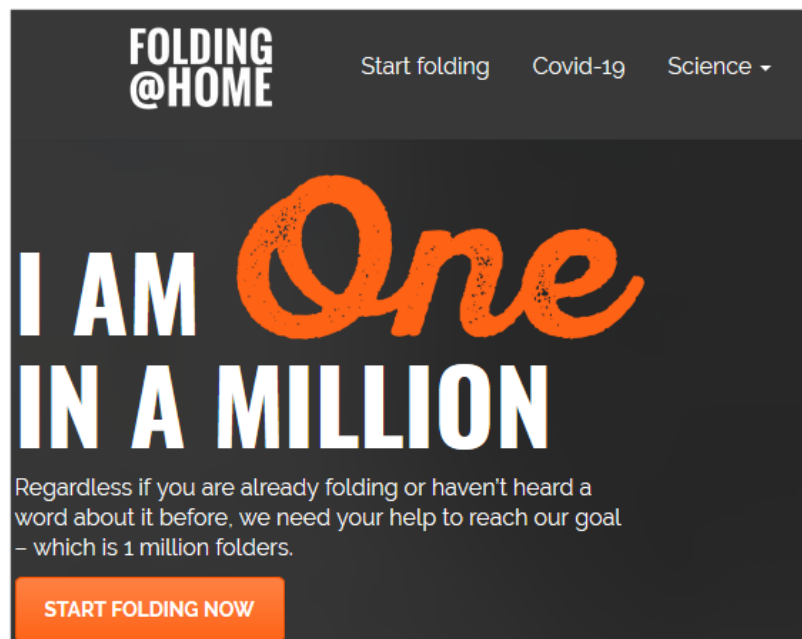
В 2007 году крупнейший проект добровольных распределенных вычислений.

После начала исследования коронавируса 2019-nCoV в проекте Folding@home

в начале 2020 года его мощность возросла с 98 PetaFLOPS до 2.7 ExaFLOPS к маю 2020 года

На тот момент лучшими суперкомпьютерами были:

- Fugaku (515 PetaFLOPS, 1st TOP500 июнь 2020)
- Summit (149 PetaFLOPS, 1st top500 ноябрь 2019)



The screenshot shows the Folding@home website with a dark background. At the top, the logo "FOLDING @HOME" is on the left, and navigation links "Start folding", "Covid-19", and "Science" are on the right. The main text reads "I AM *One* IN A MILLION" in large, bold letters, with "One" in orange script. Below this, a message states: "Regardless if you are already folding or haven't heard a word about it before, we need your help to reach our goal - which is 1 million folders." At the bottom, there is an orange button that says "START FOLDING NOW".

Top10 (top500.org*)

#	Name	Country	$R_{\max}$ (PFlop/s)
1	El Capitan	USA	1 742
2	Frontier	USA	1 353
3	Aurora	USA	1 012
4	Eagle	USA	561
5	HPC6	USA	478
6	Supercomputer Fugaku	Japan	442
7	Alps	Switzerland	435
8	LUMI	Finland	380
9	Leonardo	Italy	241
10	Tuolumne	USA	208

* Рейтинг 500 самых мощных суперкомпьютеров

Российские проекты ДРВ

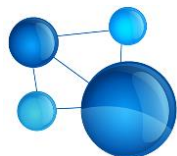
- RakeSearch (поиск латинских квадратов с опр. свойствами)
- SiDock@home (поиск лекарств от COVID-19)
совместный проект (Россия-Словения)
- USPEX@home (поиск новых материалов)
- SAT@home (криптография, SAT-подход)
- Amicable Numbers (поиск дружественных чисел)
- Optima@home (решение задач конечномерной оптимизации)
(зонтичный проект)
- Parlea@home (распределенное глубокое обучение)
- Gerasim@home (комбинаторика, латинские квадраты)
- NetMax@home (имитационное моделирование телекоммуникационных сетей)
- Acoustics@home (исследование дна Японского моря)
- AndersonAttack@home (криптография)
- XANSONS for COD (материаловедение)
- ODLK@home (поиск канонических форм латинских квадратов)

Особенности организации расчетов с помощью добровольных распределенных вычислений

- Необходимость взаимодействия с добровольцами;
- Общение на форуме;
- Научно-популярное объяснение задач проекта;
- Своевременное обновление информации на сайте проекта;
- Обеспечение достаточного количества заданий для постоянной работы проекта;
- Большой объем результатов и входных данных;
- Необходимость обработки полной серии результатов;
- Минимизация вероятности ошибочного расчета подзадачи;
- Необходимость начальной репликации заданий в связи ненадежностью узлов.

Мотивация добровольцев

- Система начисления баллов за вычисления
- Рейтинги добровольцев/команд/компьютеров
- Монеты криптовалют могут добываться в некоторых проектах:
 - Gridcoin;
 - Curecoin;
 - Foldingcoins



Curecoin (CURE)

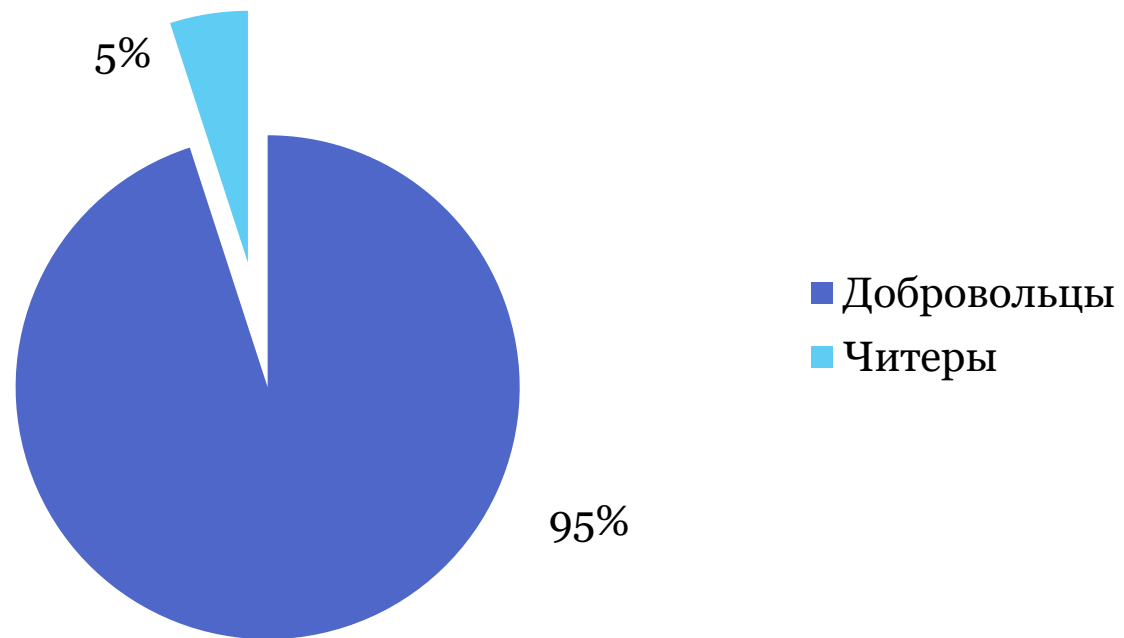


Краудсорсинговые платформы

- Расширение области применения краудсорсинга
- Проблема надежности исполнителей
- Сложность тестирования в реальных условиях

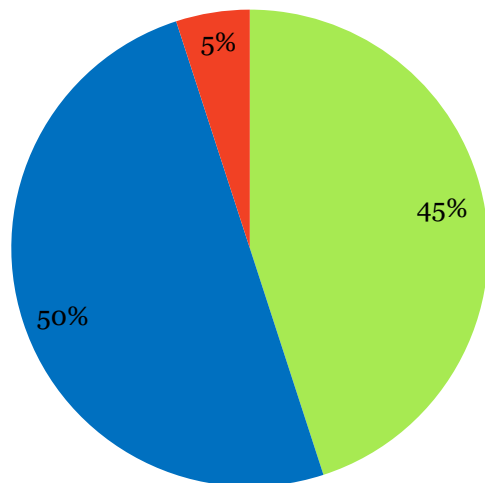


Распределение пользователей проекта ДРВ

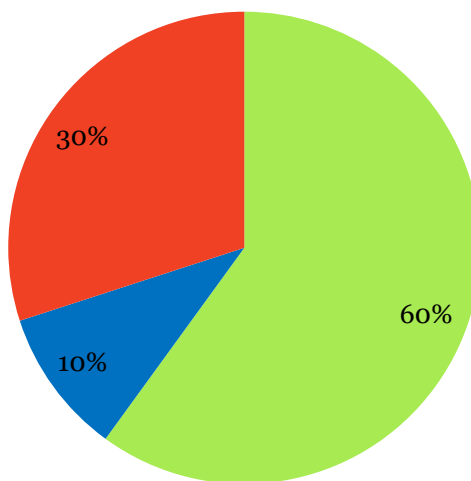


Распределения пользователей краудсорсинговой платформы

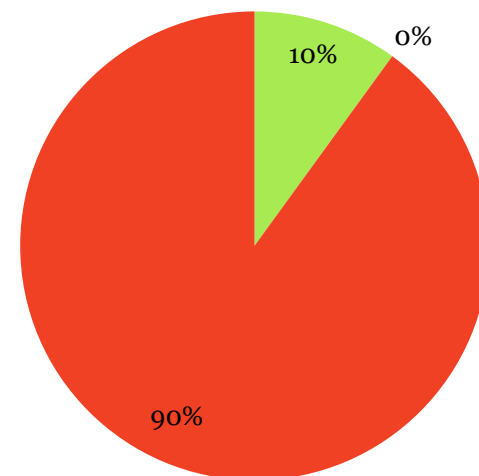
■ Новички
■ Опытные исполнители
■ Фродеры



■ Новички
■ Опытные исполнители
■ Фродеры

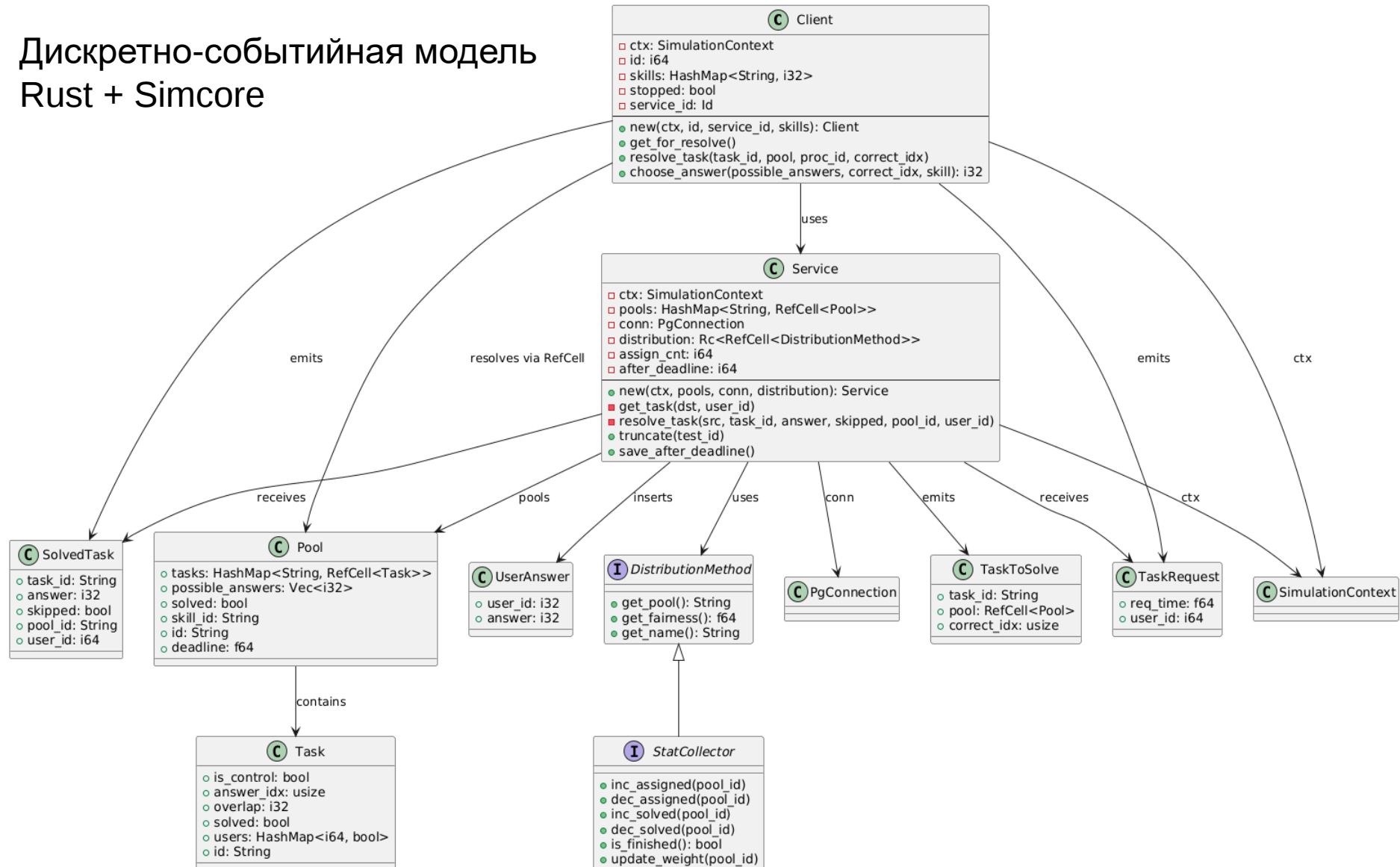


■ Новички
■ Опытные исполнители
■ Фродеры



UML-схема имитационной модели

Дискретно-событийная модель
Rust + Simcore



Алгоритмы распределения заданий

- Round Robin (RR),
- First In First Out (FIFO),
- Shortest Job First (SJF),
- **Earliest Deadline First (EDF),**
- **Fair Scheduling (FS),**
- **Deadline-Aware Fair Scheduling (DAFS),**
- **Simplified Deadline-Aware Fair Scheduling (SDAFS).**

Сценарии моделирования для алгоритмов распределения задач

- Минимальная нагрузка
- Малое количество пользователей
- Ограниченные дедлайны
- Большой разброс дедлайнов
- Нормальная нагрузка

метрика $WM(\alpha)$ – взвешенное среднее между количеством задач, решенных после дедлайна и суммарным временем простаивания пулов (наборов задач)

α – коэффициент важности просроченных задач по сравнению с временем простаивания пулов

$$WM = \alpha \cdot DL_{med} + (1 - \alpha) \cdot T_{med} ,$$

$$DL_{med} = \frac{DL - DL_{min}}{DL_{max} - DL_{min}} ,$$

$$T_{med} = \frac{T - T_{min}}{T_{max} - T_{min}} ,$$

где

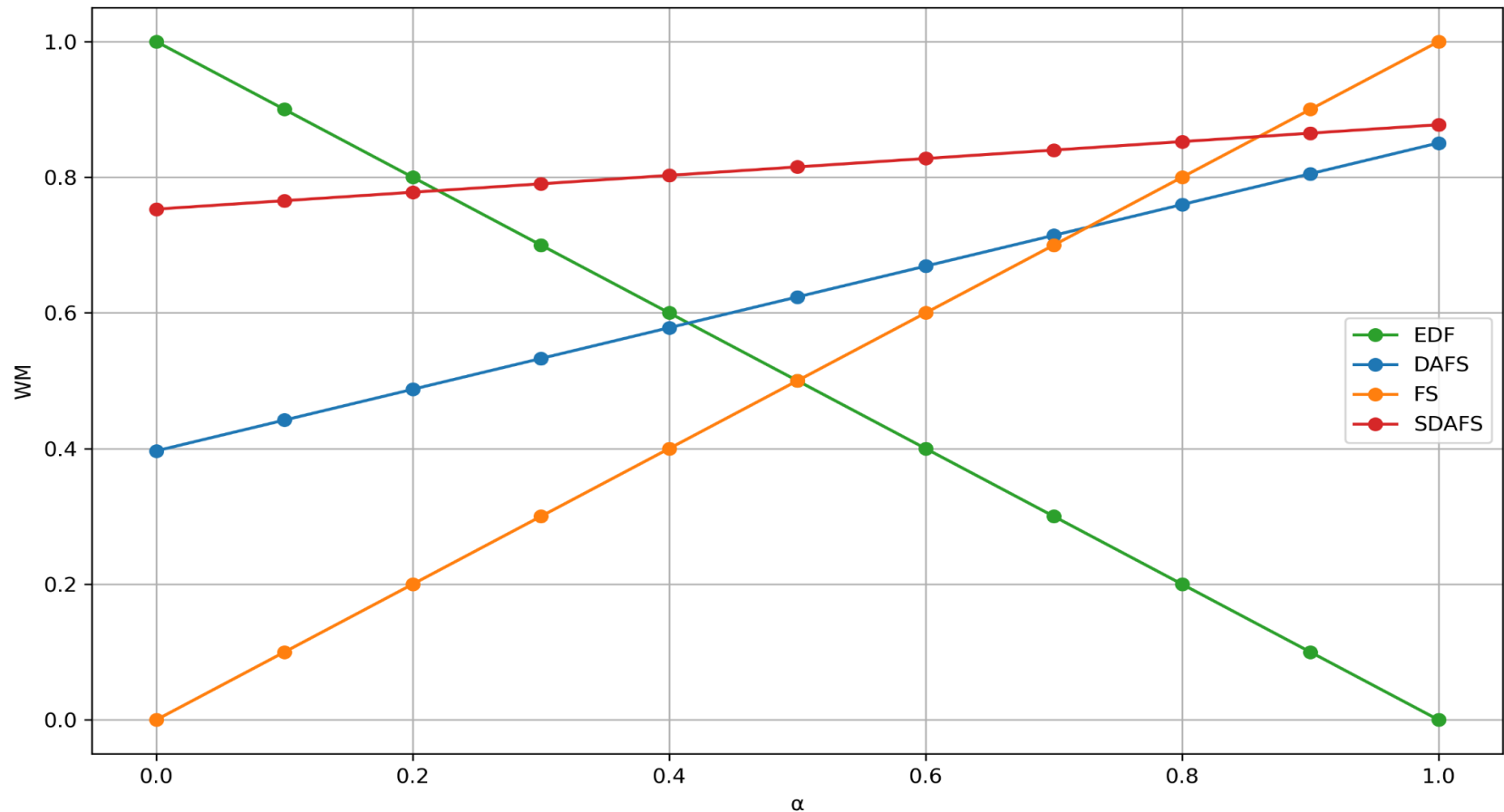
DL_{med} – нормализованное значение количества задач, решенных после дедлайна,

DL_{min}, DL_{max} – минимальное и максимальное количество задач, решенных после дедлайна по всем алгоритмам распределения задач,

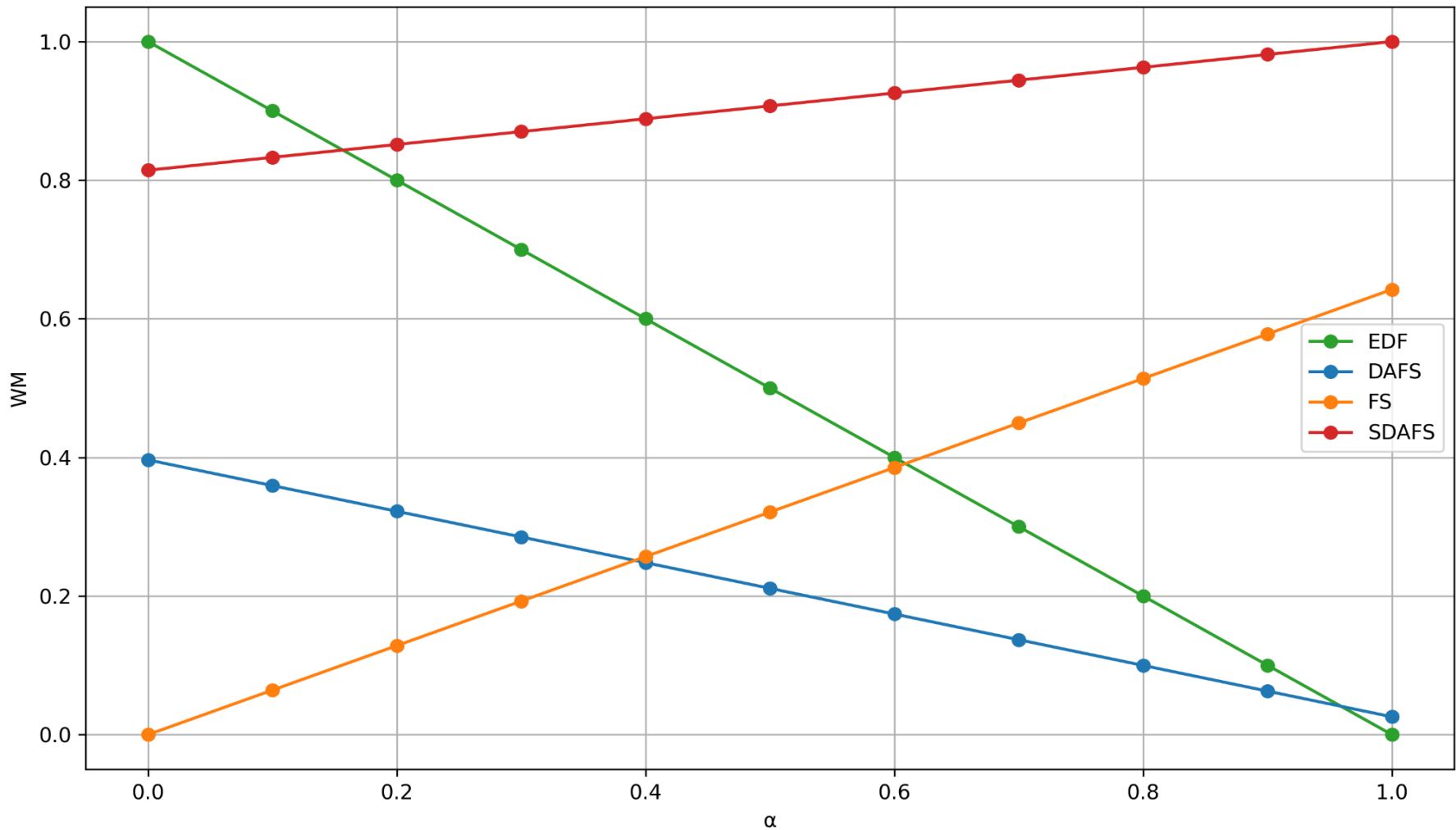
DL – количество задач, решенных после дедлайна при применении рассматриваемого метода распределения задач.

аналогичны обозначения для суммарного времени простаивания пулов T

Сценарий «Ограниченные дедлайны»



Сценарий «Большой разброс дедлайнов»



Результаты имитационного моделирования по алгоритмам распределения заданий, WM при $\alpha=0.5$

Сценарий	FS	DAFS	SDAFS	EDF
Минимальная нагрузка	0.0	0.18	0.51	0.51
Малое количество пользователей	0.19	1.0	0.36	0.5
Ограниченные дедлайны	0.0	0.35	0.36	1.0
Большой разброс дедлайнов	0.5	0.21	0.51	0.49
Нормальная нагрузка	0.5	0.35	0.4	0.5

Агрегация результатов (ответов)

Методы агрегации ответов играют важную роль в обеспечении качества и надежности решения задач на краудсорсинг-платформе

Этапы агрегации:

1. Сбор ответов;
2. Оценка качества и веса ответов;
3. Выбор финальных ответов.

Возможно наличие контрольных заданий

Методы агрегации результатов

- Majority Voting (простое большинство),
- Honeypot (вес зависит от доли верных контрольных заданий),
- Dawid-Skene, (вероятность верного ответа на полном множестве ответов)
- Модификация Dawid-Skene: One Coin, (вероятность верного ответа для последовательного представления ответов)
- GLAD (Generative model of Labels, Abilities and Difficulties)
(квалификация исполнителя и статистика по контрольным заданиям).

Результаты имитационного моделирования по методам агрегации ответов, F1-мера

Сценарий	DS	OneCoin	Majority	Honeypot	GLAD
Стандартный набор исполнителей	0.788	0.771	0.76	0.55	0.816
Стандартный набор исполнителей, контрольные задания	0.80	0.79	0.77	0.891	0.831
Низкое качество исполнителей	0.51	0.54	0.52	0.399	0.472
Низкое качество исполнителей, контрольные задания	0.475	0.478	0.44	0.415	0.479
Много фродеров	0.176	0.181	0.255	0.268	0.139
Только фродеры	0.167	0.19	0.178	0.231	0.16

Заключение

- ✓ Проведено имитационное моделирование краудсорсинговой системы с различными сценариями работы и осуществлен анализ результатов для различных методов агрегации ответов и алгоритмов распределения задач
- ✓ Разработано ПО для симуляции различных сценариев работы краудсорсинг-платформы при помощи имитационного моделирования
- ✓ Методы Majority Voting и OneCoin DS демонстрируют большую устойчивость к фроду,
- ✓ Однако при наличии большого количества контрольных заданий лучшую точность показывают GLAD и Honeyrot
- ✓ DAFS показывает лучшие результаты в условиях вариативности дедлайнов и стандартной нагрузки
- ✓ Fair Scheduling стабильно демонстрирует хорошие результаты при минимальной или ограниченной нагрузке и скорых дедлайнах, благодаря равномерному распределению ресурсов
- ✓ **Однако универсального лидера нет — выбор алгоритма должен основываться на конкретных условиях работы системы**



Спасибо за внимание

Центр распределенных вычислений ИППИ РАН

web: distributed-computing.ru

e-mail: qurochkin@gmail.com, kurochkin@iitp.ru