

Суперкомпьютерные дни в России-2026,
Москва, 28-29 сентября 2026

Distributed deep learning with adaptive local dataset formation on a heterogeneous distributed computing system

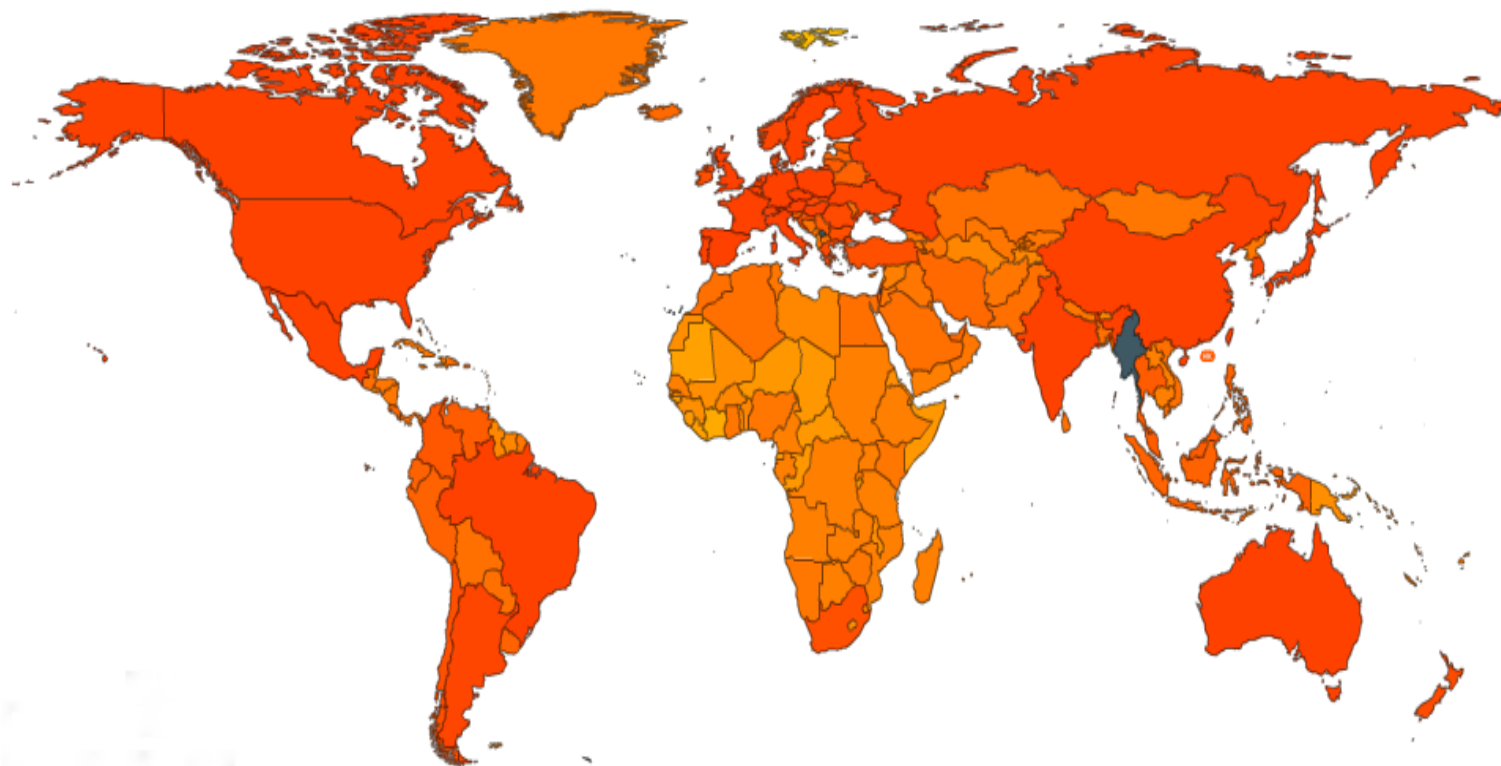
Распределенное глубокое обучение
с адаптивным формированием локального датасета
на гетерогенной распределенной вычислительной системе

И.И. Курочкин¹², В.М.Гороховатский¹

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

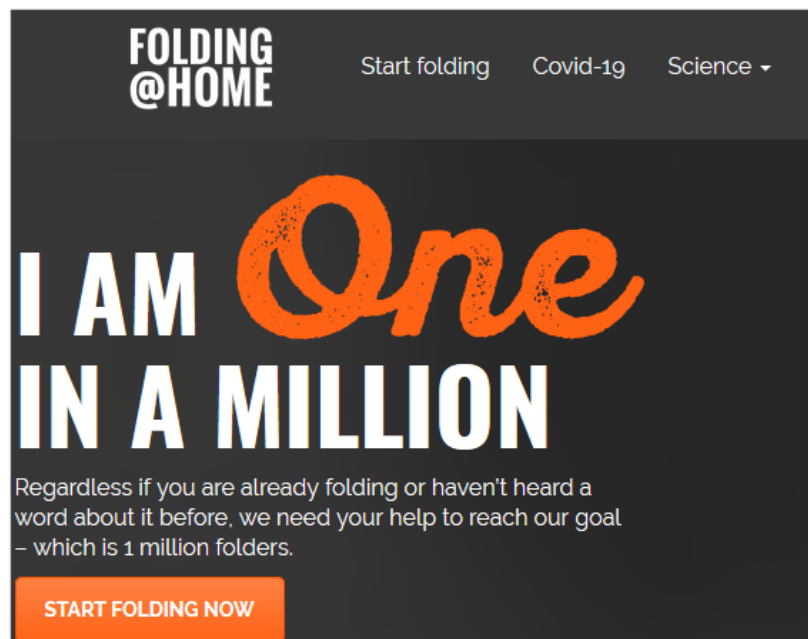
в начале 2020 года его мощность возросла с 98 PetaFLOPS до 2.7 ExaFLOPS к маю 2020 года

На тот момент лучшими суперкомпьютерами были:

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

На данный момент:

- Lineshine(2.2 ExaFLOPS, 1st TOP500 июнь 2026)



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.

Rosetta@home

3 часа назад • Мария Азарова

1 194

Нобелевскую премию по химии — 2024 присудили за вычислительный дизайн белков и предсказание их структуры 5.4

Лауреатами Нобелевской премии по химии в этом году стали Дэвид Бейкер, Демис Хассабис и Джон М. Джемпер, которые «взломали код удивительных структур белков».

Химия

AlphaFold

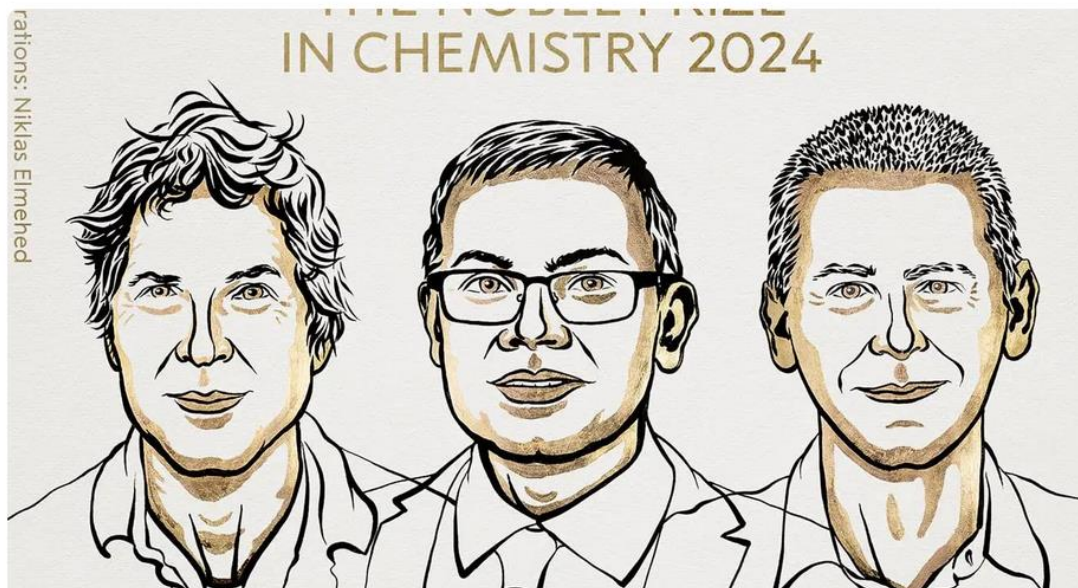
белки

награждение

Нобелевская неделя

Нобелевская премия

Нобелевская премия по химии



Анализ изображений

Задачи для обучения глубоких нейронных сетей:

- Классификация
- Сегментация
- Детекция

Ускорение обучения при использовании многопроцессорных вычислительных систем и распределенных вычислительных систем.

Необходимость использования методов распределенного глубокого обучения.

Датасеты для обучения

- CIFAR-100 (20 метаклассов, $20 * 5 = 100$ классов, 60 000 изображений 32×32)
- Lozinsky texture dataset (от 3 до 24 классов, до 4 500 000 изображений 256×256)
- Tiny Imagenet (200 классов, 120 000 изображений 240×240)
- **Imagenet ILSVRC** (200 классов, 100 000 изображений 64×64)



Виды распределенного обучения

Классификация 1:

- разделение по модели,
- **разделение по данным.**

Классификация 2 (по наличию параметрического сервера):

- **с параметрическим сервером,**
- децентрализованное.

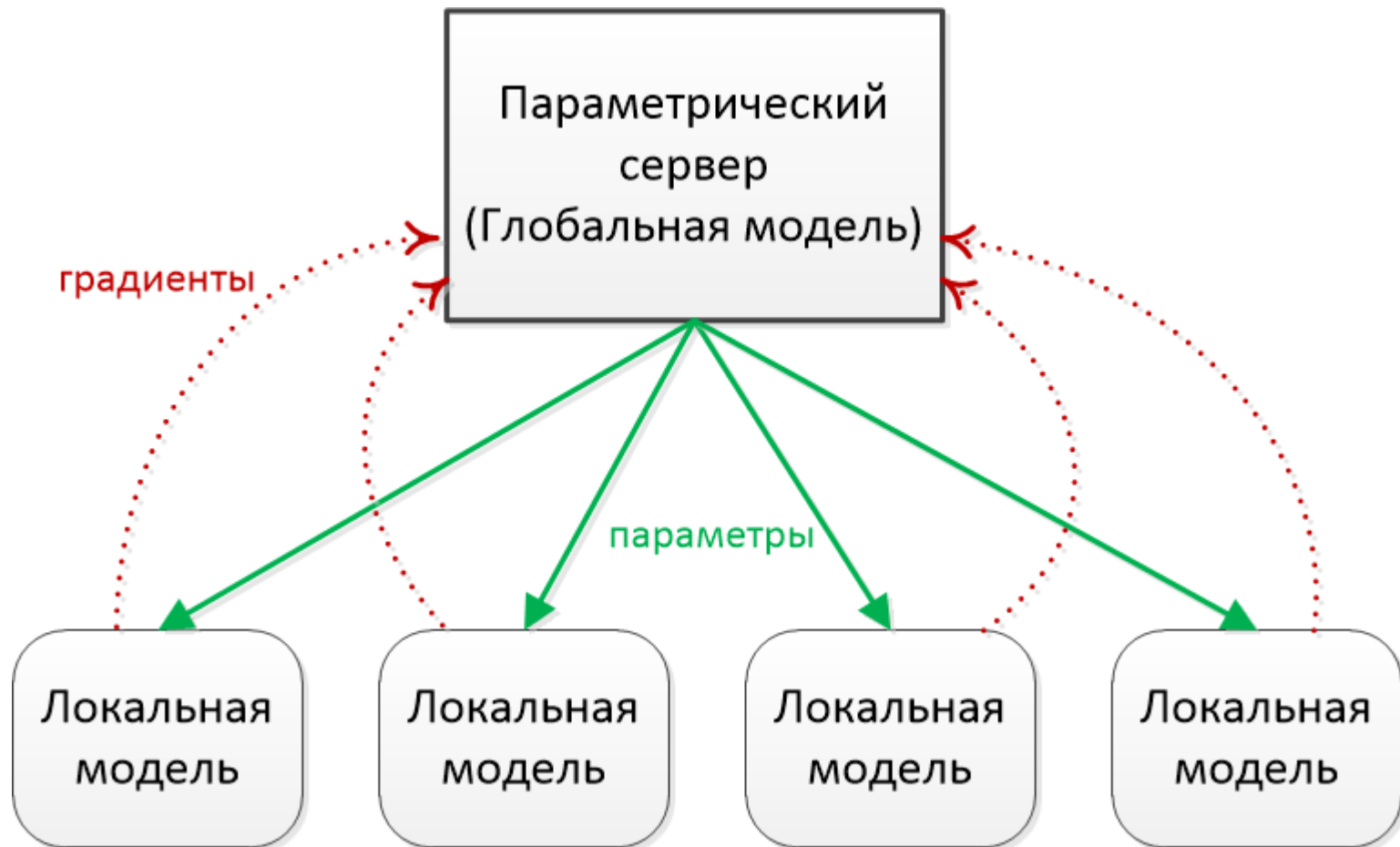
Классификация 3 (по методу агрегации данных):

- **синхронная агрегация,**
- асинхронная агрегация.

Классификация 4 (по ограничениям распространения данных):

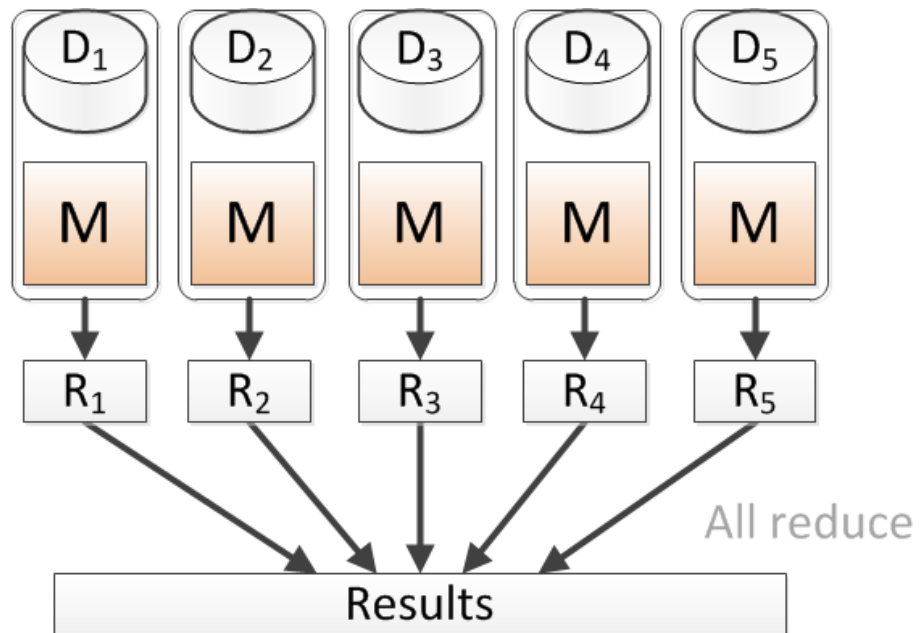
- **с общими данными,**
- федеративное обучение:
 - Горизонтальное федеративное обучение;
 - Вертикальное федеративное обучение;
 - Федеративное «трансферное» обучение.
- комбинированные данные.

Распределенное обучение с параметрическим сервером. Классификация 2

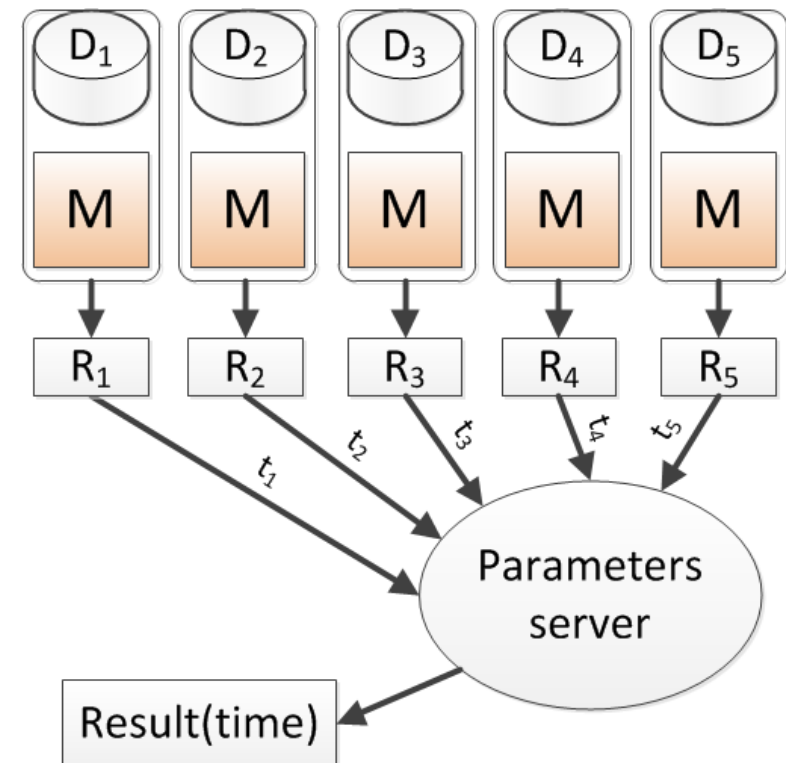


Разделение по данным

Синхронный подход



Асинхронный подход

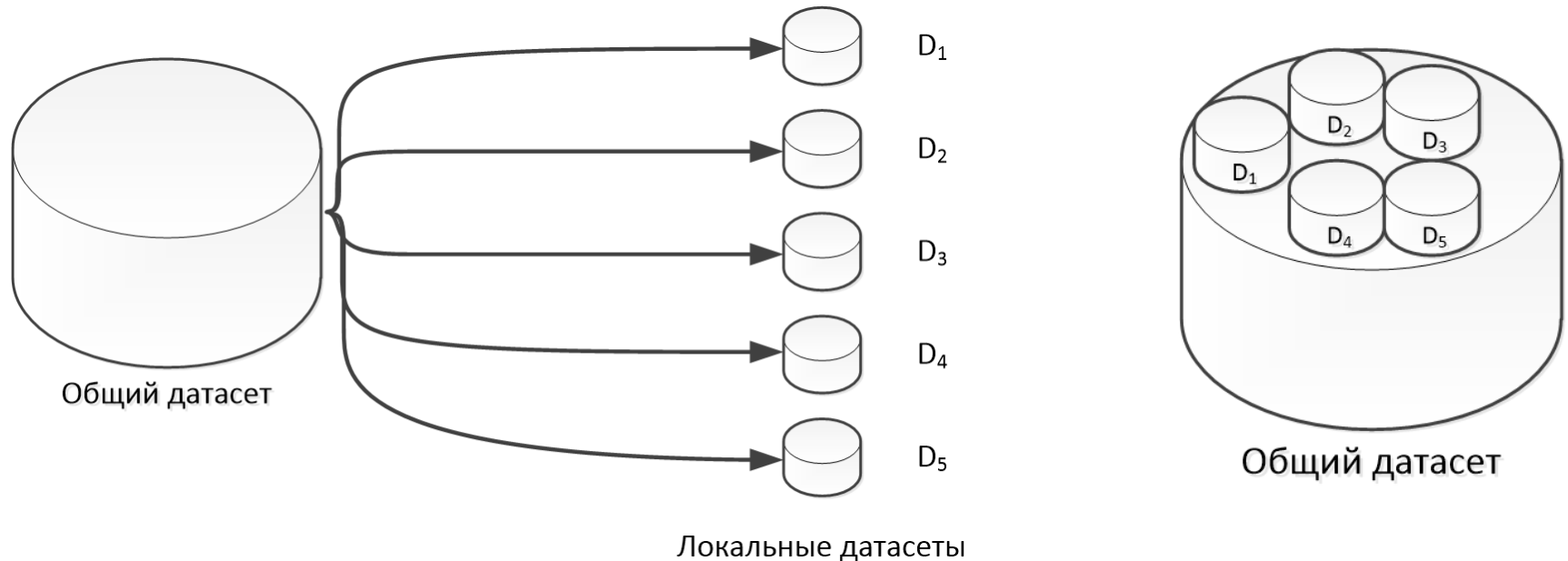


Общие данные

Локальные датасеты формируются на каждой итерации распределенного обучения.

Размер локального датасета может быть существенно меньше общего датасета.

Использование всех объектов из общего датасета достигается за множество итераций распределенного глубокого обучения



Формирование локальных датасетов

- Стохастическое (случайное) формирование локального датасета;
- Разделение датасета на части (статическое) до конца обучения;
- Разделение датасета на части (динамическое) после каждой итерации;
- Гибридное формирование датасета с использованием, как случайного выбора, так наличия части изображений из локального датасета прошлой итерации.

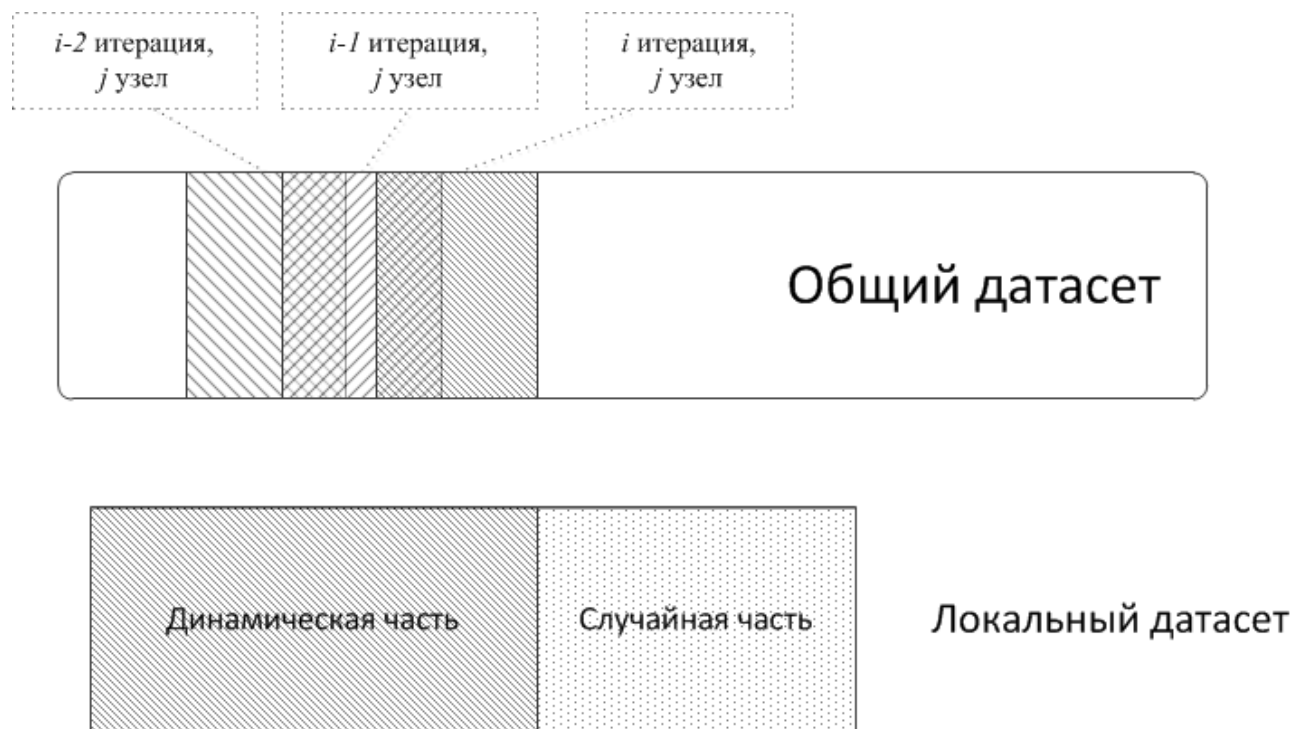
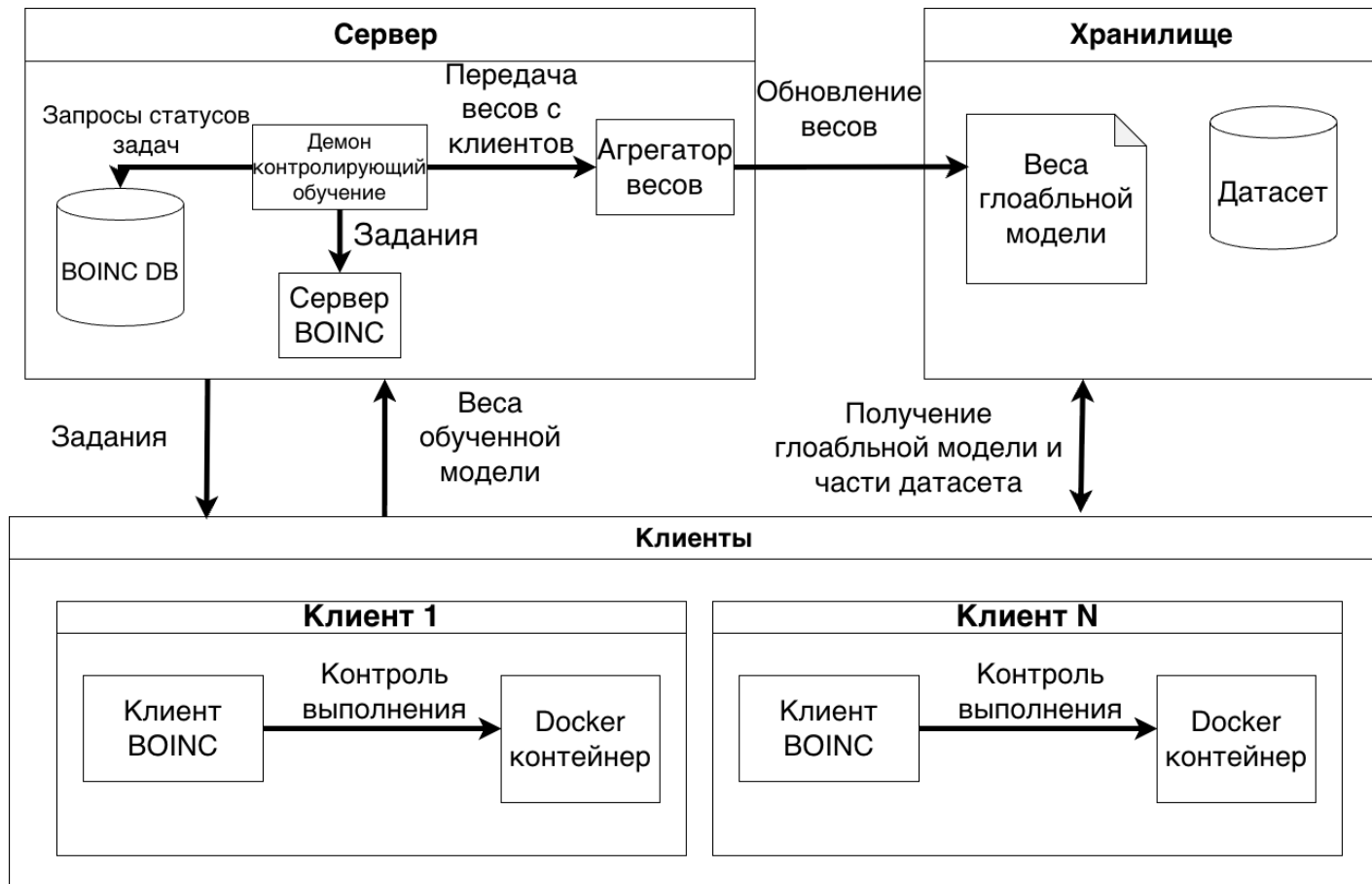
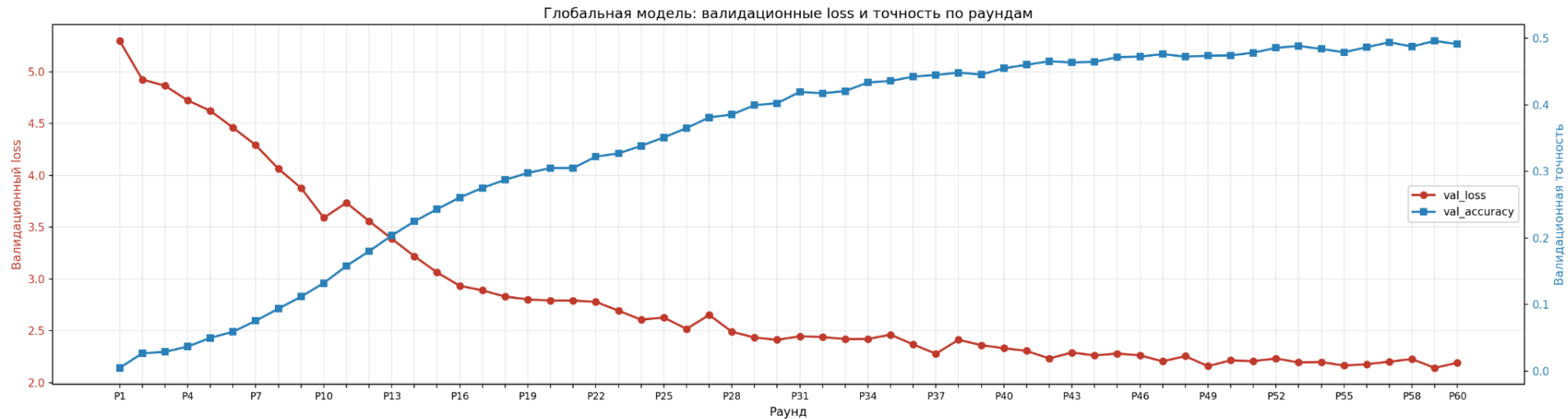


Схема распределенного обучения

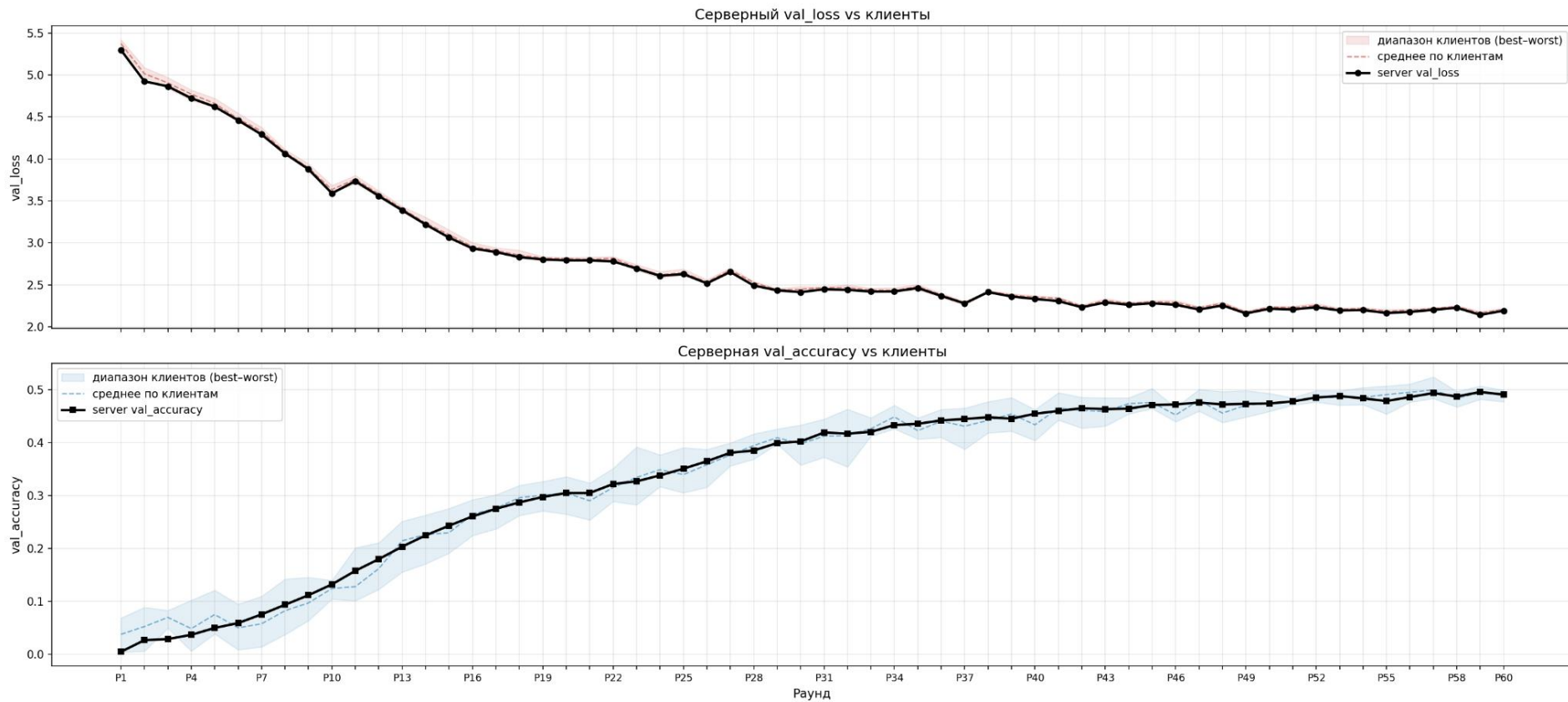


Результаты. Эксперимент 1

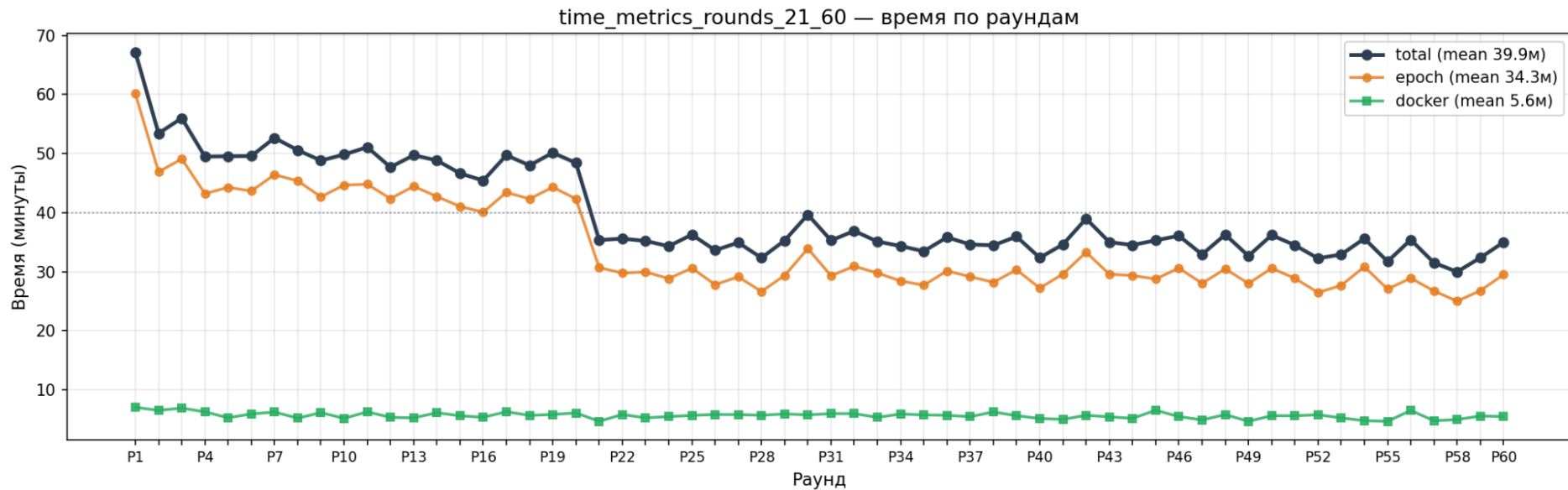
Параметр	Значение
Количество клиентов	7-12
Количество раундов распределенного обучения	60
Количество локальных эпох	5
Коэффициент обучения	0.001
Размер батча	32
Относительный размер локального датасета	4%



Динамика обучения. Эксперимент 1

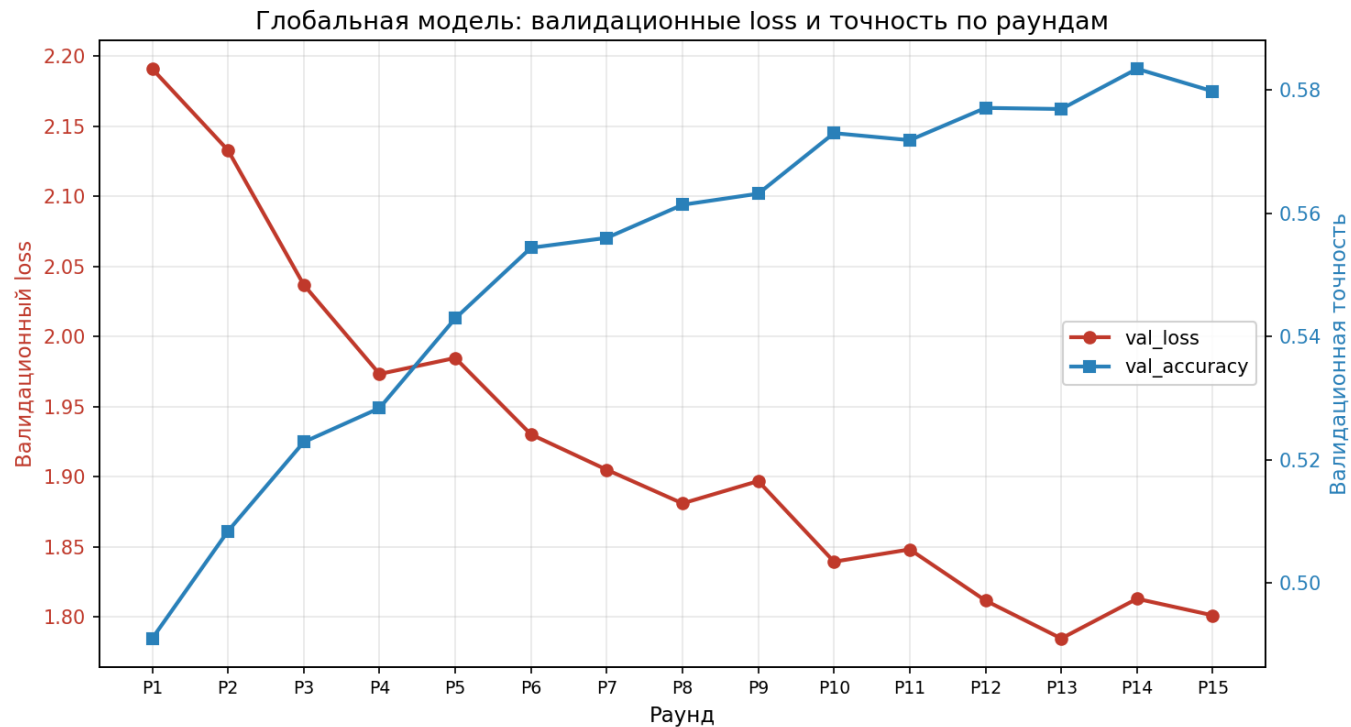


Оценка времени выполнения

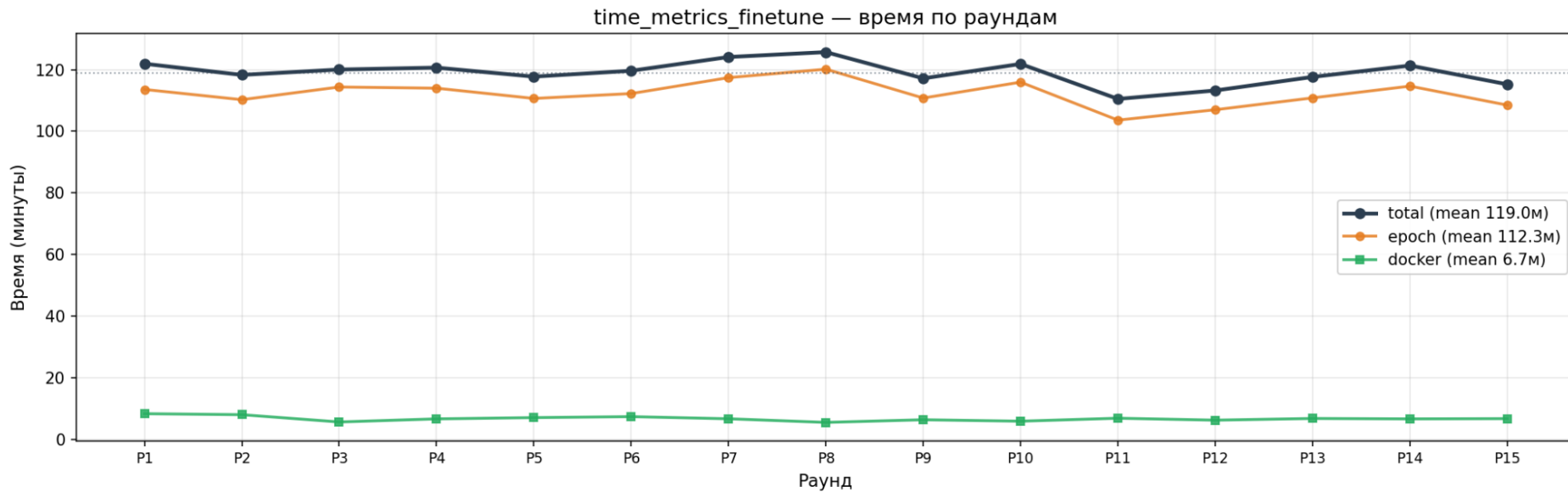


Результаты. Эксперимент 2

Параметр	Значение
Количество клиентов	7-12
Количество раундов распределенного обучения	60
Количество локальных эпох	5
Коэффициент обучения	0.001
Размер батча	32
Относительный размер локального датасета	15%



Оценка времени выполнения



Заключение

- ✓ Развернута система распределённого глубокого обучения свёрточных нейронных сетей на базе открытой платформы добровольных вычислений BOINC с использованием технологий контейнеризации
- ✓ Показаны преимущества масштабируемой инфраструктуры BOINC и системы контейнеризации Docker
- ✓ Проведены эксперименты по синхронному распределённому обучению на задаче классификации изображений
- ✓ Показаны преимущества метода формирования локальных датасетов
- ✓ Показана эффективная работа гетерогенной распределенной вычислительной системы



Спасибо за внимание

Центр распределенных вычислений ИППИ РАН

web: distributed-computing.ru

e-mail: qurochkin@gmail.com