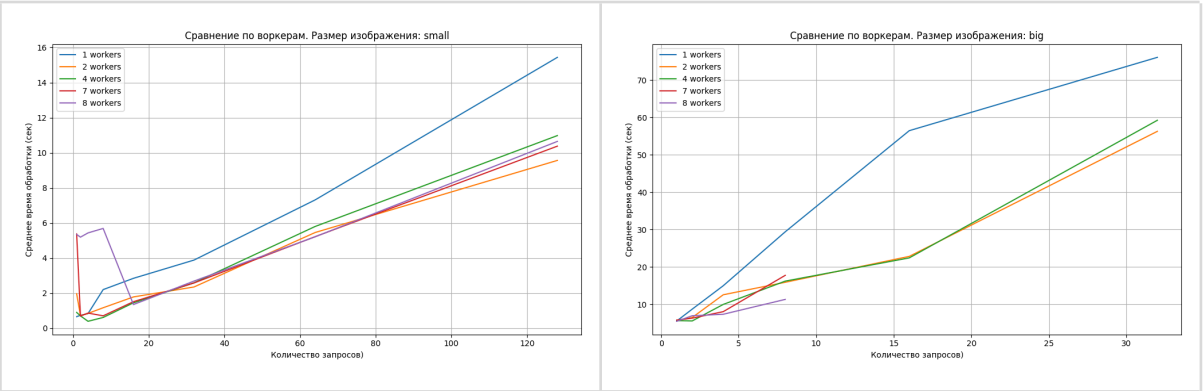
Тестовые данные

Тестируемое приложение развернуто в Yandex Cloud на виртуальной машине с параметрами: 4 ядра CPU Intel Ice Lake, 8 GB RAM, 200 GB SSD
Объем диска максимально доступный для SSD без запроса квот (чем больше размер диска, тем с большей скоростью можно читать и писать с него)
Тесты с 1 воркером были выполнены на диске 100 GB, лимиты на чтение/запись были ниже (60 mb/s вместо 108 mb/s у 200 GB)

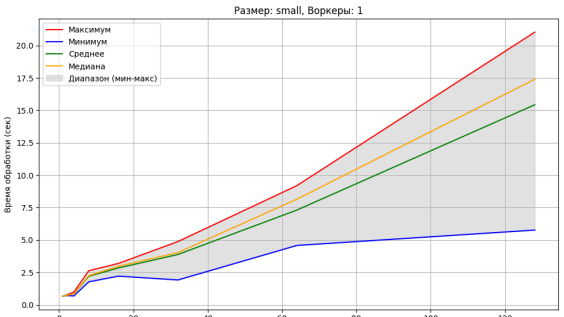
Для нагрузочного тестирования использовались 2 типа картинок. Большие (вес 50 МБ) и маленькие(143 КБ).

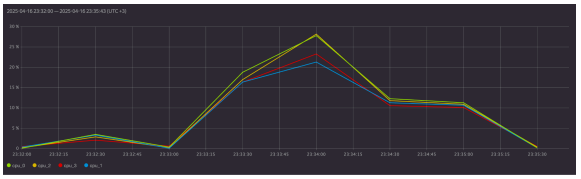
Они загружались в систему батчами по 1, 2, 4, 8, 32, 64, 128 изображений. На графиках ниже будет представлено среднее время обработки запросов в зависимости от числа изображений в нём. Цветными линиями указано разное число воркеров-обработчиков изображений.
Так же, для каждой группы тестов (1, 2, 4 воркера) будет представлен график потребления CPU и квоты диска в процентах

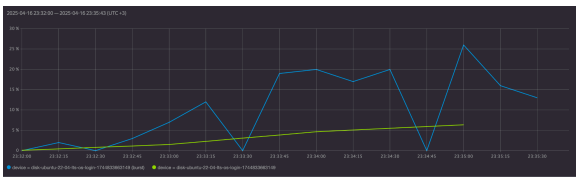
Сравнение среднего времени для разного числа воркеров

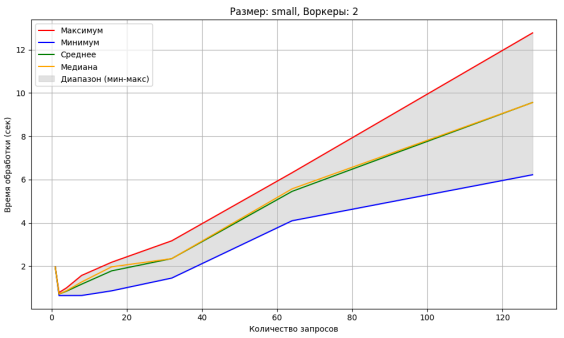


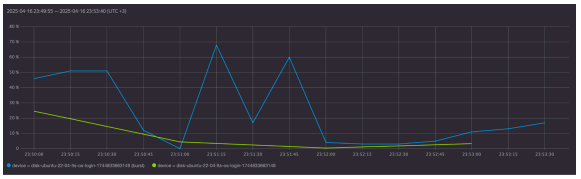
Отчет по тестированию с большом количеством легких запросов при фиксированном числе воркеров

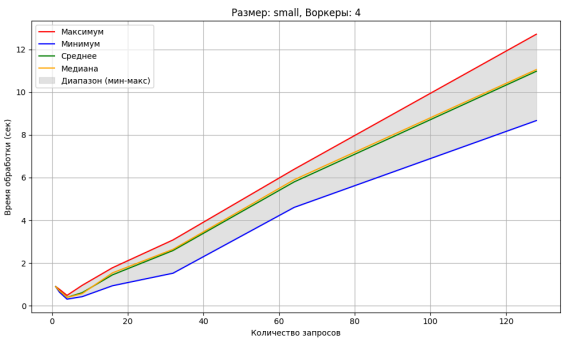


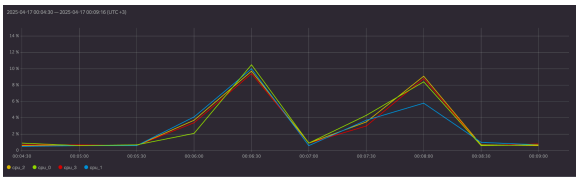


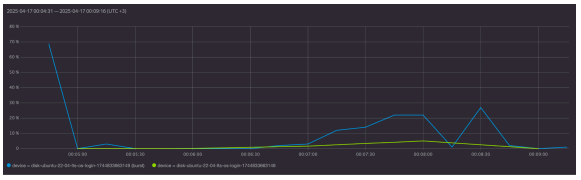


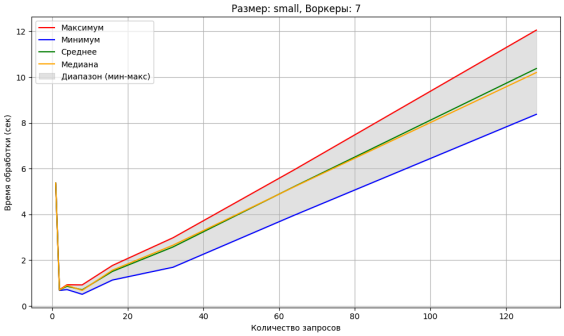


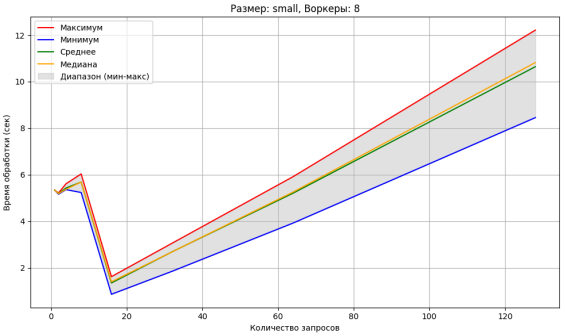




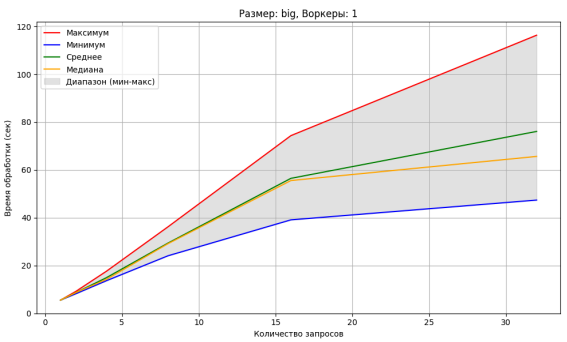


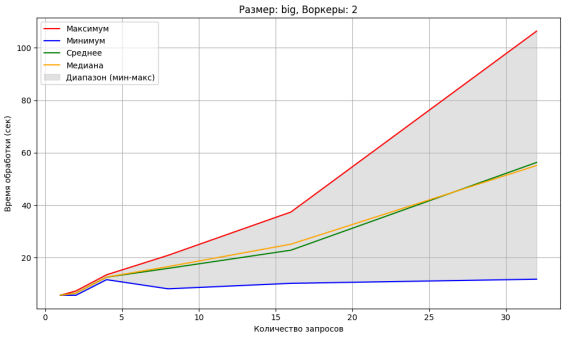


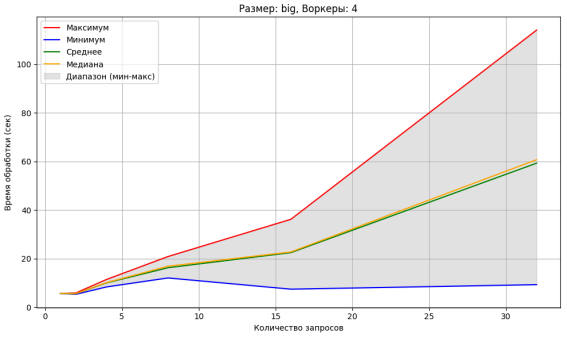


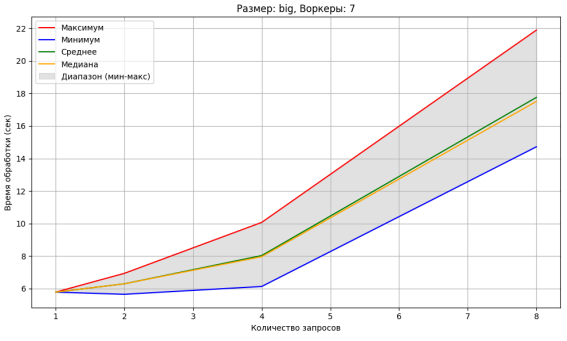


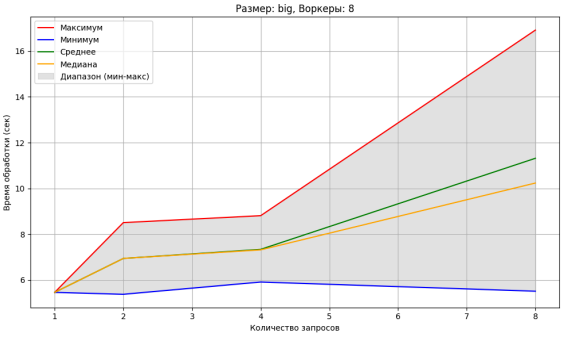
Отчет по тестированию с тяжелыми запросами при фиксированном числе воркеров











Выводы

На графиках видно, что динамика роста производительности сильно замедляется (после добавления более чем двух воркеров) из-за ограничений по квотам на запись/чтение. Также выполнение тяжелых задач на большом количестве воркеров требует много оперативной памяти, поэтому тестирование с тяжеловесными запросами без изменений ресурсов, потребляемых вм, не удалось провести полностью.

Нагрузочное тестирование позволило сделать следующие заключения:

1) В текущем решении самое узкое место — скорость чтения/записи на диск
2) Для большего числа воркеров необходимо больше ОЗУ (1 воркер потребляет где-то 1 Гб ОЗУ).
   Следовательно могут быть следующие варианты решения:
   - Увеличение объема ОЗУ, чтобы больше загружать процессор и эффективнее пользоваться ресурсами
   - Распределение воркеров на разные ноды, чтобы увеличить объем ОЗУ для каждого воркера
   - Использовать паралеллельность, встроенную в celery. Это позволит экономичнее использовать ОЗУ, но выполнять те же задачи за то же время.
3) S3 ограничен скоростью дисковых операций, что требует либо масштабирования (распределённого развертывания для равномерного распределения нагрузки), либо использования более производительных дисков (SSD/NVMe)
4) Стоит ограничить максимальный объём файлов/объём одного изображения, загружаемых в приложение на 1 запрос, чтобы не вызвать Out Of Memory