

Инструкция для экспертной проверки программного обеспечения

1. Подготовка к установке

1.1 Установка должна производиться на устройства отвечающие следующим минимальным требованиям к аппаратному обеспечению:

- ЦП: x86_64 архитектура, от 2 ядер
- ОЗУ: от 4 ГБ (зависит от размера используемой языковой модели).
- GPU: Nvidia или AMD
- Дисковая подсистема: SSD от 25 ГБ (зависит от размера загруженной языковой модели).
- ОС: **Рекомендуется Astra Linux, ОСнова.**
Поддержка:
 - Дистрибутив Linux с поддержкой python 3.8+ (Debian, Ubuntu),
 - Windows 10/11

1.2 До установки основного ПО убедитесь что установлено внешнее ПО:

- ПО источник запросов, например Open WebUI, если отсутствует, можно отправлять прямые запросы к эндпоинту клиента в консоли (в формате OpenAI API)

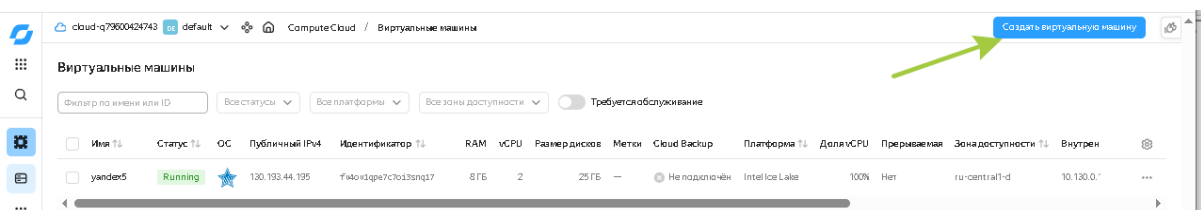
- LLM сервер, например Ollama или llama.cpp (поддерживающий работу с эндпоинтом в формате OpenAI API или Ollama API) и установленными моделями

1.3 Возможный вариант развертывания виртуальной машины

Создание VM на примере Yandex Cloud, приведенный в видеоинструкции по установке

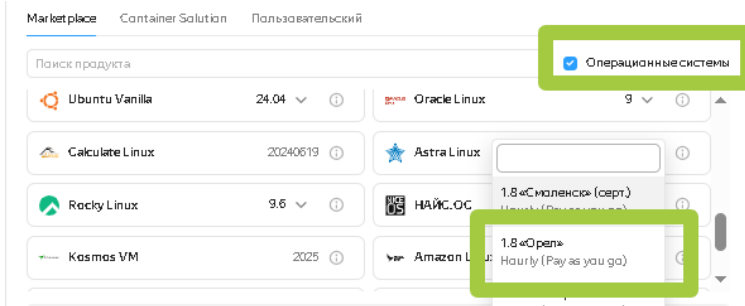
Варианты создания инфраструктуры не ограничиваются ПО

1. Выберите раздел Все сервисы - compute cloud, нажмите на кнопку создания VM

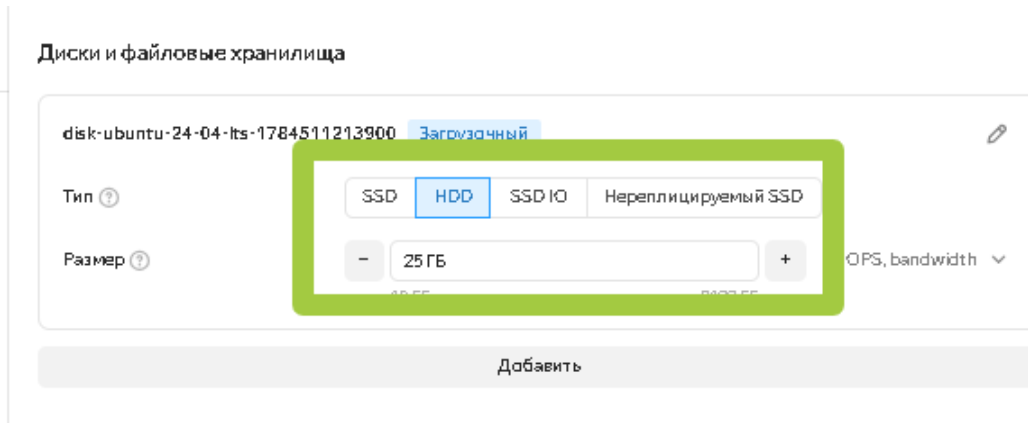


2. В разделе выбора ОС установите чекбокс Операционные системы, найдите Astra

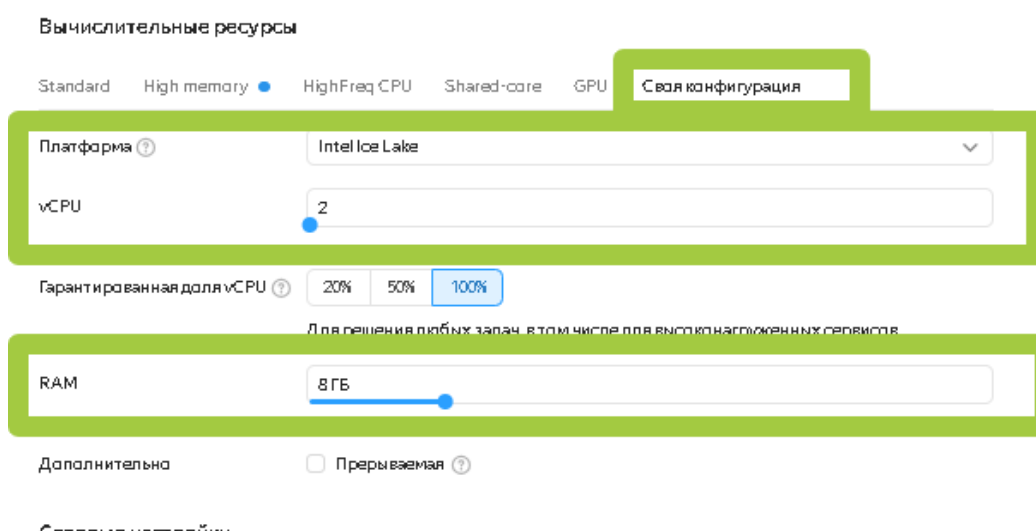
linux и выберите в выпадающем списке версию 1.8 "Орел"



3. Установите размер HDD или SSD на менее 25 Гб



4. В вкладке Своя конфигурация установите платформу, количество vCPU и RAM в соответствии с минимальной конфигурацией



5. Установите параметры доступа

Доступ

SSH-ключ ?
Данные для аутентификации хранятся на ВМ и не могут быть изменены. В случае утери закрытого SSH-ключа доступ к ВМ станет невозможен.

Доступ по OS Login ?
Более безопасное подключение с помощью SSH-сертификатов или SSH-ключей. Легко восстановить доступ даже в случае утери ключа.

Логин* ?
user

SSH-ключ* ?
ssh-key-1784247559527 X v или **Добавить ключ**

Резервное копирование ? ☐

После создания ВМ, следуйте инструкции по установке

2. Установка ПО

2.1 Установка ПО на Astra linux 1.8 «Орел»

#Создание папки ПО

```
sudo mkdir -p /opt/pbdr
```

#Установка прав для пользователя

```
sudo chown -R $USER:$USER /opt/pbdr
```

```
cd /opt/pbdr
```

Загрузите ПО (файл архива) в рабочую папку /opt/rag/ и распакуйте содержимое вручную или с помощью команд

```
bash
```

```
curl -L -o pbdr.zip https://oqoai.ru/pbdr.zip
```

```
sudo apt install -y unzip
```

```
unzip -o pbdr.zip
```

Повторите процесс установки для всех устройств используемых при тестировании

Необходимо развернуть не менее 1 PBDR сервера на устройстве с установленным LLM сервером (Ollama, llama.cpp и др.), не менее 1 PBDR клиента на устройстве с установленным LLM сервером (Open WebUI и др.), 1 PBDR сервер администрирования на любом устройстве с браузером

Все устройства на которые разворачивается ПО должны находиться в 1 сети и иметь статические IP адреса

Внимание! Рекомендуется использовать 1 PBDR сервер с ollama, 1 PBDR сервер с llama.cpp, 1 PBDR клиент для тестов маршрутизации

[Далее к разделу первый запуск](#)

3. Первый запуск

Порядок первого запуска:

1. Запуск PBDR сервера

1.1. Запустите PBDR серверы на устройствах используемых в качестве серверов

Linux:

```
chmod +x ./pbdr_server_OAA4  
./pbdr_server_OAA4 pbdr_server_config_test.json
```

Windows:

```
pbdr_server_OAA4.exe pbdr_server_config_test.json
```

2. Запуск PBDR клиента

2.1 Сконфигурируйте IP адреса LLM серверов в конфигурации клиента. Укажите IP адреса и порты узлов на котором размещены PBDR сервера

Настройка групп

При необходимости объедините созданные узлы в группы

Действия администрирования

В панели управления при необходимости:

- Используйте функции Обновить статус
- Используйте функции Добавить устройство
- Используйте функции Добавит группу
- Используйте функции Метрики GPU

В панели узлов при необходимости:

- Используйте фильтрацию по узлам - клиентам или узлам - серверам
- Выполните обновление статуса
- Отредактируйте информацию о узле
- Просмотрите метрики узла - сервера
- Откройте окно редактирования конфигурации узла
- Откройте окно редактирования политик узла - клиента
- Удалите ненужные узлы

5.Тестовые запросы

Для тестирования используйте Внешнее ПО для отправки запросов.

Для OpenWebUI в разделе Панель администратора - Настройки -Подключения

Укажите эндпоинт запросов конечного пользователя в формате:

http://<IP хоста клиента>:<порт хоста>/v1

например:

http://192.168.1.105:8080/v1

В списке моделей OpenWebUI появятся модели доступные для запроса

Выполните запрос к LLM модели. Запрос будет маршрутизирован к узлу который оптимален с точки зрения политик и PBDR алгоритмов, ответ на запрос вернется после обработки LLM сервером

Поддержка

По вопросам, связанным установкой и эксплуатацией ПО, обращайтесь в техподдержку:

Телефон: +79600424743

E-mail: q79600424743@yandex.ru