

ОПИСАНИЕ И УСЛОВИЯ ПРЕДОСТАВЛЕНИЯ УСЛУГ  
УСЛУГА «AI Cloud - Inference»

## 1. НАИМЕНОВАНИЕ УСЛУГИ

- 1.1. Наименование Услуги: «AI Cloud - Inference».
- 1.2. Настоящий документ содержит описание состава Услуги, ее базовой функциональности, возможных сопутствующих и дополнительных услуг, общего порядка подключения, изменения и отключения Услуги, условий предоставления и ограничений.

## 2. ИНФОРМАЦИЯ ОБ УСЛУГЕ

## 2.1. Краткое описание Услуги

Услуга предоставляет веб-сервис для эффективной сборки Docker-образов на базе моделей Машинного и Глубокого обучения, а также для их дальнейшего разворачивания на ресурсах SberCloud в виде микросервисов со сгенерированным API.

Для оказания Услуги Заказчику необходимым условием является наличие на его площадке подключения к сети интернет, достаточного для эффективной загрузки данных, моделей или их производных (например, чекпоинтов моделей или сериализованных моделей) на сервер.

С помощью Услуги Заказчик может разворачивать модели искусственного интеллекта на базе инфраструктуры SberCloud и сервиса AI Cloud - Inference для дальнейшего обращения к ним посредством API-запросов от АС Заказчика.

Заказчику для успешной реализации вывода моделей искусственного интеллекта в виде микросервисов предоставляется возможность сборки образа с любым Программным обеспечением, python-библиотеками и способом взаимодействия с моделями искусственного интеллекта.

На Рисунке 1 приведена общая упрощенная схема взаимодействия с сервисом AI Cloud - Inference с удаленной площадки Заказчика (с указанием зон ответственности):

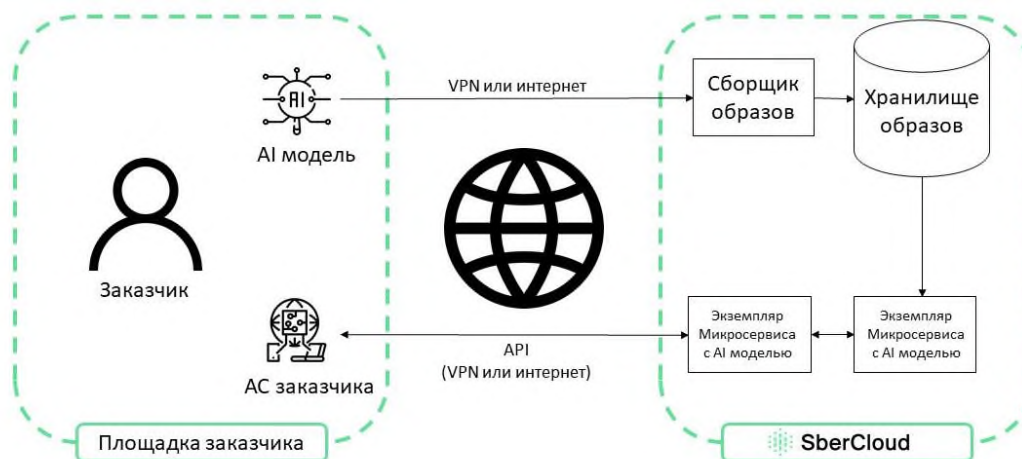


Рисунок 1: Схема взаимодействия Заказчика с Сервисом разворачивания моделей машинного и глубокого обучения на мощностях кластера AI Cloud - Inference.

В зоне ответственности SberCloud - функционирование Серверов с развернутыми моделями искусственного интеллекта, функционирование вычислительного кластера, Объектного хранилища S3, а также прочей инфраструктуры.

В рамках Услуги Заказчик может самостоятельно отслеживать и управлять состоянием развернутых моделей.

Для подключения к Услуге Заказчик может выбрать один или несколько типов подключения:

- Подключение через общий канал Интернет (shared) предполагает логическое подключение к общему для всех Заказчиков Услуги каналу передачи данных. Скорость сетевого соединения для каждого Заказчика не является гарантированной и зависит от загруженности общего канала передачи данных (Услуга предоставляется по умолчанию).
- Подключение через прямой канал связи. Данный способ подключения позволяет обеспечить взаимодействие сетей Заказчика с сетью в облаке с помощью выделенных каналов связи стороннего провайдера. Опционально, с помощью данного сценария, к Услуге Заказчика может быть подключен альтернативный канал в сеть Интернет. Для данного подключения могут быть использованы выделенные каналы Заказчика, организованные с использованием «темной оптики» (Услуга оплачивается отдельно).

Услуга предоставляется на базе защищенной инфраструктуры облачной платформы «AI Cloud», меры защиты которой приведены в описании услуги AI Cloud - Model Training.

## 2.2. Условия хранения данных в Объектном хранилище S3

Хранение, использование и тарификация хранения и использования данных в Объектном хранилище S3 осуществляется в рамках соответствующей Услуги по Договору.

## 2.3. Использование сервиса AI Cloud - Inference

Создание, конфигурация и разворачивание моделей искусственного интеллекта осуществляется напрямую Заказчиком.

## 2.4. Техническое описание решения

### 2.4.1. Программная платформа

Услуга реализуется средствами веб интерфейса, внутреннего docker registry, сборщиком образов и комплексом KFServing/Knative/ISTIO/Kubernetes. Посредством них и программных библиотек пользователь имеет возможность собирать и разворачивать модели искусственного интеллекта в виде микросервисов.

### 2.4.2. Аппаратная платформа

Вычисления и обсчет задач осуществляется на предоставляемой Заказчику в рамках Услуги области кластера Кристофари, а также других вычислительных ресурсов и серверов SberCloud.

### 2.4.3 Технические особенности и ограничения:

Скорость загрузки данных на площадку Исполнителя ограничена пропускной способностью канала доступа в Интернет из инфраструктуры Заказчика до облака SberCloud, а также скоростью чтения данных с СХД Исполнителя.

Общие значения параметров услуги «AI Cloud - Inference»

| Описание  | Мин. значение | Макс. значение   |
|---|---------------|--|
| Размер предоставляемого хранилища S3                    | 1 Гб          | 20 Тб  |
| Количество утилизируемых в рамках вычисления задачи GPU | 1 GPU         | В соответствии с количеством свободных GPU, отображаемом в Личном кабинете |

## 3. ТАРИФИКАЦИЯ УСЛУГИ

### 3.1. Возможные виды тарификации Услуги:

#### 3.1.1. Динамическая тарификация (Pay as you go).

### 3.2. Стоимость Услуги формируется в зависимости от количества GPU/CPU, на которых происходило вычисление запросов к API микросервисов с моделями, а также самого времени, в течении которого вычислялись запросы

к API микросервисов с моделями искусственного интеллекта и объема зарезервированного Заказчиком Объектного хранилища S3.

- 3.3. Момент начала списания денежных средств – с момента начала вычисления запроса к API микросервиса (для каждого запроса).
- 3.4. Момент окончания списания денежных средств – с момента окончания вычисления запроса (для каждого запроса).

#### **4. ИНЫЕ УСЛОВИЯ, ПРИМЕНИМЫЕ К УСЛУГЕ**

- 4.1. Возможные виды подключения / изменения / отключения Услуги:
  - 4.1.1. Посредством подписания Заказа;
  - 4.1.2. Посредством совершения действий на Портале.
- 4.2. Возможный порядок расчётов по Услуге:
  - 4.2.1. Предварительная оплата;
  - 4.2.2. Постоплата (на основании отдельно заключенного письменного бланка Заказа).
- 4.3. Возможные способы оплаты / порядок пополнения баланса:
  - 4.3.1. Оплата в безналичном порядке на основании выставленного Исполнителем счёта;
  - 4.3.2. Оплата посредством электронных средств платежа.
- 4.4. В связи с характером потребления Услуги, а также объёмов, которых она может достигнуть в Отчётный период, Стороны установили, что в случае заключения с Заказчиком соглашения о применении Постоплаты (пп. 4.2.2. настоящего Приложения) Заказчик выбирает лимит в пределах Отчётного периода, по достижении которого Услуги оказываются на основании соответствующего обращения уполномоченного лица в Контактный Центр и, по требованию Исполнителя, предоставления гарантийного письма.