https://habr.com/ru/companies/oxygendc/news/806843/

vvroschin

11.04.2024

Серверы с GPU все чаще арендуют для тренировок ИИ, а за год операторы заработали на них 6,6 млрд рублей

Серверы с графическими процессорами (GPU) для «тяжелых» технических задач пользуются все большим спросом на российском рынке облачных услуг. Так, доля аренды таких серверов для обучения моделей искусственного интеллекта (ИИ) в 2023 году составила 5,4%. Об этом сообщают «Ведомости» со ссылкой на директора по развитию бизнеса iKS-Consulting Дмитрия Горкавенко.

Собеседник издания уточнил, что общая выручка операторов от аренды таких серверов в прошлом году составила почти 6,6 млрд руб. При этом в 2022 году на услугу аренды серверов с GPU приходилось 4%, или 3,7 млрд руб. То есть выручка выросла на 44%.

Еще 4,8%, (5,9 млрд руб.), в 2023 г. пришлось на услуги вычислений на суперкомпьютерах. Годом ранее этот сервис дал в 2,5 раза меньше дохода – 2,4 млрд руб., или 2,6% от общей выручки.

Совокупный же объем российского рынка облачных инфраструктурных сервисов (IaaS и PaaS) в 2023 году – 121,4 млрд руб., что на 25% больше, чем в 2022 году, когда общий доход сегмента составил 90,6 млрд руб.

По прогнозам iKS-Consulting, к 2030 году доля выручки от услуги по аренде серверов с GPU увеличится почти вдвое до 8,6%, или до 50,3 млрд руб., при общем объеме рынка в 585,1 млрд руб.

Серверы с GPU применяются, в частности, для обучения генеративных моделей, распознавания и синтеза речи, работы цифровых ассистентов или распознавания лиц в видеопотоке. Но задачи для таких серверов могут быть самыми разными. Наши клиенты, например, работают на серверах с GPU с графикой и видео, организуют VDI, транслируют и конвертируют «тяжелые» видео, запускают на них особенно требовательные к производительности программы для рендеринга, 3D-графики и моделирования.

«Железо» для таких целей, разумеется, нужно мощное. У нас это ускорители Tesla® V100s, Tesla® A40 и Tesla® A100 с двойной точностью (double precision). Достоверность и надежность вычислений такие ускорители обеспечивают на 2-3% больше, чем у популярных моделей NVidia® GTX 3080. А тренировка нейросетей происходит в 8 раз быстрее, чем на «железе» с CPU. Больше о облачных сервисах OXYGEN можно узнать по этой ссылке.

А вы работаете с GPU-серверами? Какие задачи решаете с их помощью? Какие модели ускорителей считаете лучшими на рынке? Расскажите об этом в комментариях!