На конференции Cloud Next 25, Google представила свой самый мощный ИИ-чип — Ironwood, разработанный специально для задач инференции (использования уже обученных моделей).

Это седьмое поколение собственных ускорителей компании и первый чип, полностью ориентированный на выполнение ИИ-задач в реальном времени.

В конфигурации из 9216 чипов Ironwood достигает невероятной производительности — 42,5 EFLOPS, что в 24 раза выше, чем у одного из самых мощных суперкомпьютеров мира — El Capitan. При этом Ironwood вдвое энергоэффективнее предыдущего TPU поколения Trillium.

Ключевые особенности Ironwood:

- 192 ГБ HBM-памяти на чип, что в 6 раз больше, чем у Trillium;
- Пропускная способность памяти — 7,2 ТБ/с, это в 4,5 раза выше предыдущего поколения;
- Интерчиповый канал связи — 1,2 Тбит/с в обе стороны, +50% к прошлому TPU.

Google предложит Ironwood в облачном сервисе Google Cloud в двух конфигурациях: на 256 чипов и на 9216 чипов — под задачи разного масштаба. Компания уверена, что именно инференция станет ключевым направлением в развитии ИИ, а Ironwood должен задать новый стандарт в этой области.