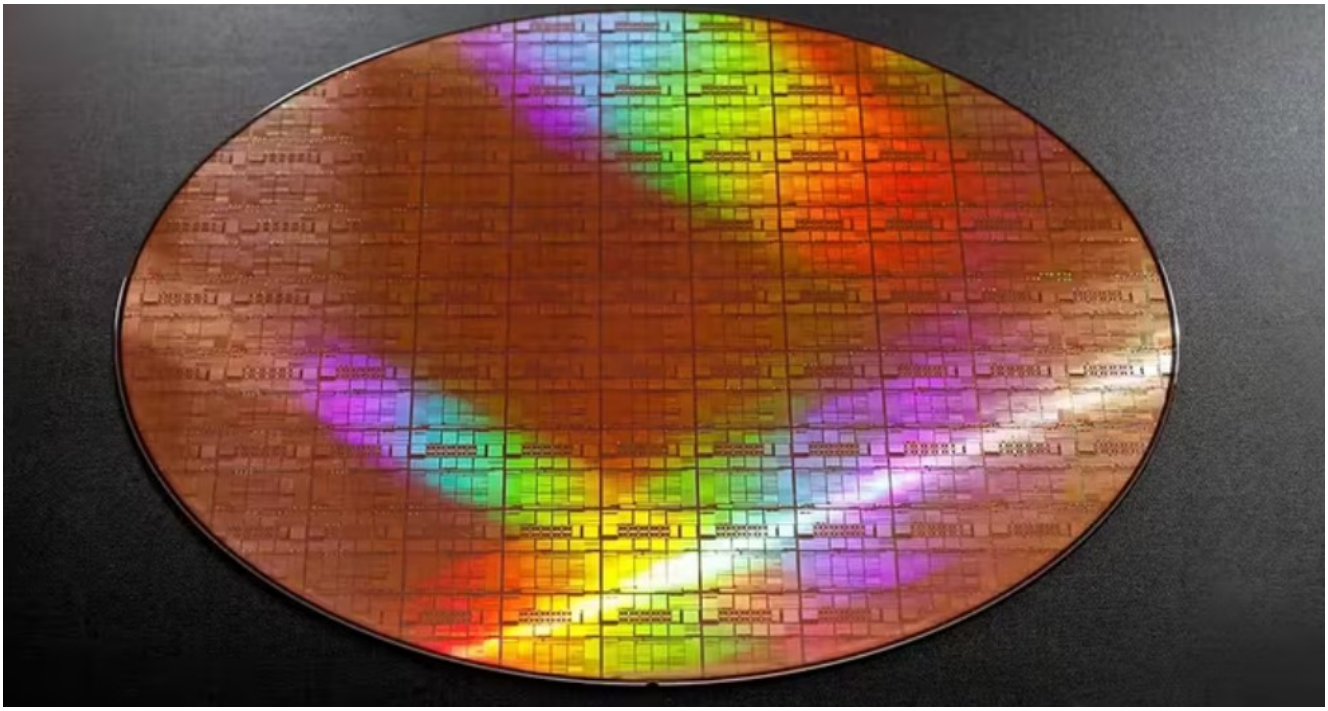
Intel y AMD añaden a la arquitectura x86 un conjunto de instrucciones eficiente para IA

Intel y AMD se han visto amenazadas en el sector de los centros de datos por la arquitectura ARM, así que crearon hace dos años el Grupo Asesor del Ecosistema x86.

Su trabajo conjunto ha fructificado en un nuevo conjunto de instrucciones eficiente para inteligencia artificial, que es el talón de Aquiles de la arquitectura x86 y que la británica Arm ha aprovechado. Al menos hasta ahora, porque la especificación de Extensiones de Cómputo para IA (ACE) ya ha sido publicado.



Estas extensiones se basan en lo ya existente de AVX10, un conjunto de instrucciones desarrolladas por Intel, pero que por el bien de la arquitectura x86 se ha usado para mejorarla y ser más atractiva para los centros de datos. Casi todos los

cálculos de inteligencia artificial tienen que ver con matrices, y sobre todo multiplicaciones con baja precisión, y eso es al final lo que han añadido Intel y AMD, pero con estructuras a nivel físico más eficientes, y que se puedan compactar más para aumentar la densidad de cómputo de los procesadores x86.

De hecho, se apoya en los registros que ya tiene AVX10 y suma ocho nuevos registros de matriz de 16×16, con los que las dos compañías afirman alcanzar hasta dieciséis veces más potencia de cálculo que con las instrucciones AVX10 para esta tarea. Es un caso maximalista, porque será complicado alcanzarlo en un uso real, pero no imposible.

Este conjunto de instrucciones está pensado sobre todo para inferencia, donde se trabaja con formatos de datos de baja precisión, así que admite de forma nativa los más habituales en IA, como INT8, BF16 o los FP8 y MXFP8 del estándar OCP. Tener *hardware* específico para multiplicar matrices con esas precisiones, en lugar de hacerlo con instrucciones más genéricas, es lo que lleva a la mejora de consumo y la eficiencia.

AMD ya ha adelantado que su arquitectura Zen 7 traerá un «nuevo motor de matrices» y una «ampliación de formatos de datos de IA», que apuntan justo a ACE; Intel lo incorporará en una futura arquitectura. En ese momento cualquier lenguaje de programación podrá hacer uso de las instrucciones (PyTorch, TensorFlow, etc.) sin tener que escribir código distinto para distinto *hardware* x86. Que total, siendo las IA las que solo van a programar para cuando lleguen los Zen 7, tampoco es un problema tan grave, pero mejora la eficiencia.

Fuente: geektopia.