A Apple juntou-se ao conselho de administração do Ultra Accelerator Link Consortium, dando-lhe mais voz sobre como a arquitetura da infraestrutura de servidores de IA irá evoluir.

O Ultra Accelerator Link Consortium (UALink) é um grupo de padrão aberto da indústria para o desenvolvimento de especificações UALink. Como potencial elemento-chave utilizado para o desenvolvimento de modelos e aceleradores de inteligência artificial, o desenvolvimento dos padrões poderia ser extremamente benéfico para o futuro da própria IA.

Na terça-feira, foi anunciado que mais três membros foram eleitos para a diretoria do consórcio. A Apple fazia parte do trio, ao lado de Alibaba e Synopsys.

O consórcio é agora composto por mais de 65 empresas como membros desde a sua constituição em outubro de 2024.

“O ULink mostra-se muito promissor na abordagem dos desafios de conectividade e na criação de novas oportunidades para expandir as capacidades e demandas de IA”, disse Becky Loop, Diretora de Arquitetura de Plataforma da Apple. “A Apple tem uma longa história de pioneirismo e colaboração em inovações que impulsionam a nossa indústria e estamos entusiasmados por nos juntarmos ao Conselho de Administração da UALink.”

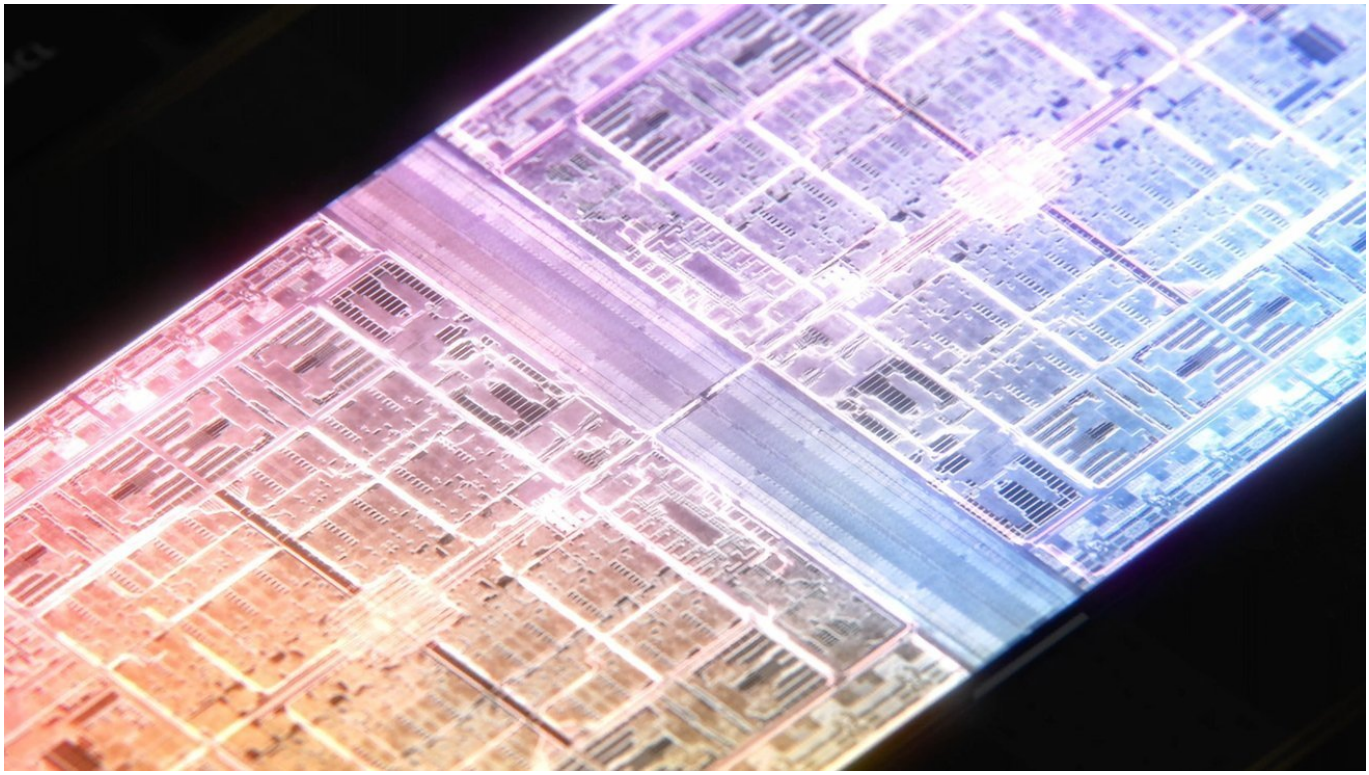
O presidente do conselho do consórcio UALink, Kurtis Bowman, deu as boas-vindas às três empresas no conselho. “O apoio contínuo ao Consórcio ajudará a acelerar a adoção deste padrão-chave da indústria, definindo a interconexão de próxima geração para cargas de trabalho de IA”, disse ele.

Interconecte-se com o futuro

O UALink é descrito como uma “tecnologia de interconexão de acelerador de alta velocidade e expansão que aprimora o desempenho do cluster de IA de próxima geração”. O consórcio se encarrega de desenvolver as especificações técnicas para as interconexões que residem entre aceleradores de IA, ou GPUs.

Em suma, as interconexões são utilizadas para realizar conectividade de alta largura de banda entre dois componentes de processamento, para minimizar quaisquer gargalos e incentivar comunicações rápidas. No caso aqui, é para permitir que várias GPUs ou chips de IA se comuniquem entre si com atraso mínimo, para que possam trabalhar juntos como se fossem um chip maior.

Isso é semelhante em conceito à interconexão que a Apple usa em seus chips Apple Silicon Ultra, para conectar dois chips Max.


A interconexão UltraFusion no M1 Ultra – Crédito da imagem: Apple

O conceito, quando se trata de servidores UALink e AI, é que a interconexão conectaria vários chips. Como descreve o UALink, “centenas de aceleradores em um pod”, com a interconexão também permitindo carregamento e armazenamento simples de semântica “com coerência de software”.

Em termos simples, o UALink prevê o uso de uma interconexão para conectar muitos chips de IA e GPUs, juntamente com comunicações extremamente rápidas entre os componentes. Tudo para que possam trabalhar mais rapidamente no desenvolvimento e processamento de IA.

Atualmente, o grupo pretende lançar a especificação UALink 1.0 no primeiro trimestre de 2025. A expectativa é permitir até 200 Gbps de largura de banda por pista, com possibilidade de conectar até 1.024 aceleradores em um pod de IA.

Um benefício futuro da Apple

Como uma empresa que avança no mundo do desenvolvimento de IA, em parte através da introdução do Apple Intelligence, a Apple tem interesse em orientar o desenvolvimento de IA.

Na verdade, existem vários aspectos em jogo que a Apple pode aproveitar como parte do conselho UALink.

O mais óbvio é o desenvolvimento de servidores com chips de IA de alto desempenho. Já

considerou utilizar vários sistemas para desenvolver os modelos de IA utilizados nos seus produtos, mas um hardware melhor pode acelerar os processos de aprendizagem ou permitir que mais processos ocorram simultaneamente.

Em última análise, isso pode economizar dinheiro em recursos ou manter os mesmos gastos, mas obter mais benefícios.

Isso não seria apenas para fins de treinamento de modelos, pois também é possível que os servidores aprimorados que usam as interconexões possam ser usados para consultas baseadas em nuvem.

A Apple tenta realizar seu processamento no dispositivo, mas também emprega servidores para consultas mais difíceis fora do dispositivo. Com servidores mais rápidos, essas consultas poderiam ser respondidas mais rapidamente, ou com mais processamento aplicado, do que atualmente.

Também pode haver um elemento relacionado ao processamento no dispositivo. Embora pretenda que um grande número de componentes se comuniquem entre si, a Apple poderia usar o que aprendeu para seu próprio hardware.

Além da interconexão dos chips Ultra, a Apple também depende muito da conectividade de alta velocidade em seus chips em geral. Otimizar o funcionamento de suas criações de sistema no chip aumentará seu desempenho, o que beneficiará mais diretamente os usuários finais.

Este último objetivo pode ser extremamente útil para chips futuros, mas não está claro se será usado no Apple Silicon ou não no momento. O uso certo no curto prazo será para hardware de servidor.

Mesmo assim, com a especificação de primeira geração chegando em meses, ainda pode demorar muito até que as interconexões se tornem mais comumente usadas no campo da IA.