Photonic computing beyond Moore's Law

Prof. Dr. Wolfram Pernice

Kirchhoff-Institut für Physik, Heidelberg

Ever noticed that annoying lag that sometimes happens during the internet streaming from, say, your favorite football game? Called latency, this brief delay between a camera capturing an event and the event being shown to viewers is surely annoying during the decisive goal at a World Cup final. But it could be deadly for a passenger of a self-driving car that detects an object on the road ahead and sends images to the cloud for processing. A way to dramatically reduce latency in artificial intelligence (AI) systems lies in using light for computation instead of electronic circuits. Combining photonic processing with what's known as the non-von Neumann, in-memory computing paradigm enables to perform computations with unprecedented, ultra-low latency and compute density. Photonic tensor cores run computations at a processing speed higher than ever before and perform key computational primitives associated with AI models such as deep neural networks for computer vision, with remarkable areal and energy efficiency. While scientists first started tinkering with photonic processors back in the 1950s, in-memory computing (IMC) is an emerging non-von Neumann compute paradigm where memory devices, organized in a computational memory unit, are used for both processing and memory. By removing the need to shuttle data around between memory and processing units, IMC even with conventional electronic memory devices could bring significant latency gains. However, the combination of photonics with IMC could further reduce the latency issue – so efficiently that photonic in-memory computing might soon play a key role in latency-critical AI applications. Together with in-memory computing, photonic processing overcomes the seemingly insurmountable barrier to the bandwidth of conventional AI computing systems based on electronic processors.