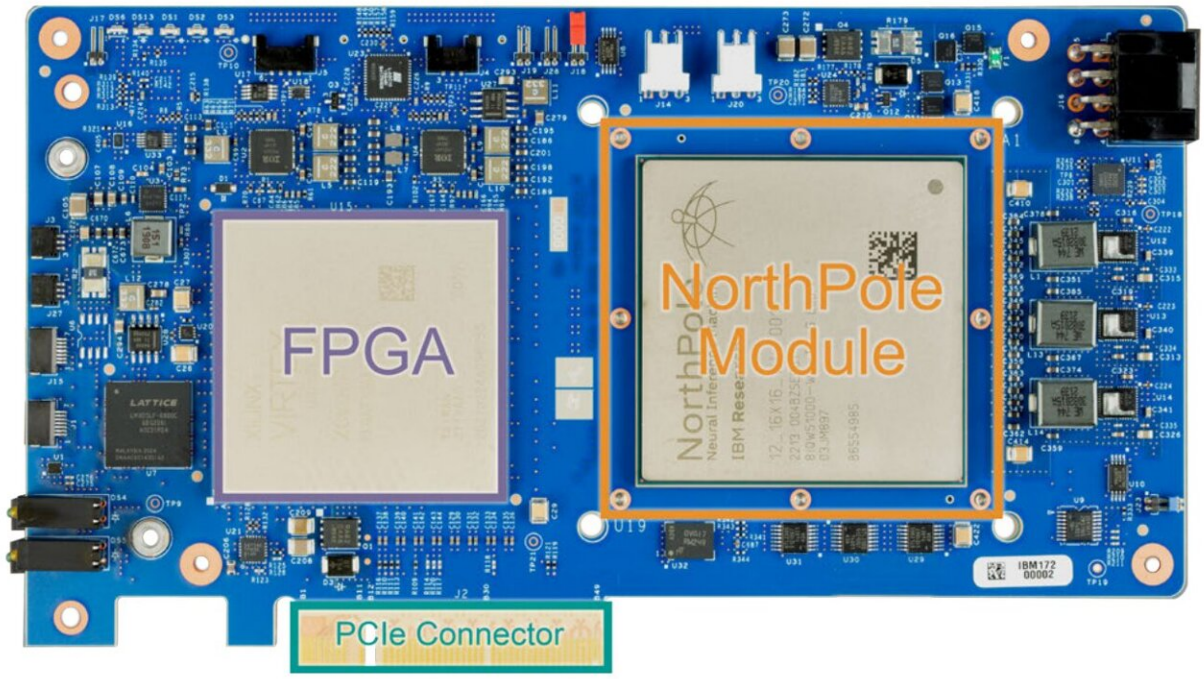
IBM's NorthPole chip runs AI-based image recognition 22 times faster than current chips

October 20 2023, by Bob Yirka



Printed circuit board used for bring-up and testing of the NorthPole chip. The packaged NorthPole module is interfaced to the board through a socket to permit testing multiple chips (socket lid 5 is not shown). Communication between the NorthPole module and a host computer is mediated by the FPGA and occurs through a ×8 PCIe edge connector. The FPGA only acts as a PCIe bridge and, in future, this functionality can be implemented on NorthPole. Note that the board has no external memory. Credit: *Science* (2023). DOI: 10.1126/science.adh1174

A large team of computer scientists and engineers at IBM Research has developed a dedicated computer chip that is able to run AI-based image recognition apps 22 times as fast as chips that are currently on the market.

In their paper published in the journal *Science*, the group describes the ideas that went into developing the chip, how it works and how well it performed when tested. Subramanian Iyer and Vwani Roychowdhury, both at the University of California, Los Angeles, have published a Perspective piece in the same journal issue, giving an in-depth analysis of the work by the team in California.

As AI-powered applications become mainstream tools used by professionals and amateurs alike, scientists continue work to make them better. One way to do that, Iyer and Roychowdhury note, is to move toward an "edge" computer system in which the data is physically closer to the AI applications that are using them.

Commercial applications such as ChatGPT, for example, rely on data accessible through the internet, which introduces time delays. In this new effort, the research team at IBM has developed and built a computer chip that combines the processing module and the data it uses—they named it NorthPole. The team reports that the design of their chip was inspired by the way the human brain works.

The chip uses a two-dimensional array of memory blocks and interconnected CPUs to accomplish its tasks—its all-digital architecture allows the computing cores to communicate directly with far-away blocks as easily as with those that are nearby, a design that allows the chip to process data and return answers quickly.

The research team ran identical applications on its chip and several others that are currently available on the market, including NVIDIA

GPUs. They found NorthPole completed tasks up to 22 times faster than any of the other chips. They also found it delivered faster transistor speeds.

The researchers acknowledge that their new chip does suffer from one major fault—it is only able to run specialized AI processes; thus, it cannot run training processes, or large language models like ChatGPT. But they also note that soon, they will test connecting multiple NorthPole chips together—a move that they believe will overcome its current limitations.

More information: Dharmendra S. Modha et al, Neural inference at the frontier of energy, space, and time, *Science* (2023). DOI: 10.1126/science.adh1174

Subramanian S. Iyer et al, AI computing reaches for the edge, *Science* (2023). DOI: 10.1126/science.adk6874