NAND RESEARCH

RESEARCH NOTE

INSIDE THE IBM NORTHPOLE NEURAL ACCELERATOR

STEVE MCDOWELL, CHIEF ANALYST
OCTOBER 30, 2023

CONTEXT

IBM Research in Almaden, California, has developed a groundbreaking AI chip called NorthPole, which could revolutionize AI hardware systems. Unlike traditional computer chips, NorthPole integrates processing units and memory on the same chip, eliminating the von Neumann bottleneck and significantly improving efficiency.

IBM NorthPole integrates processing units and memory on a single chip, which promises to significantly improve energy efficiency and processing speed for artificial intelligence tasks. It is designed for low-precision operations, making it suitable for a wide range of AI applications while eliminating the need for bulky cooling systems.

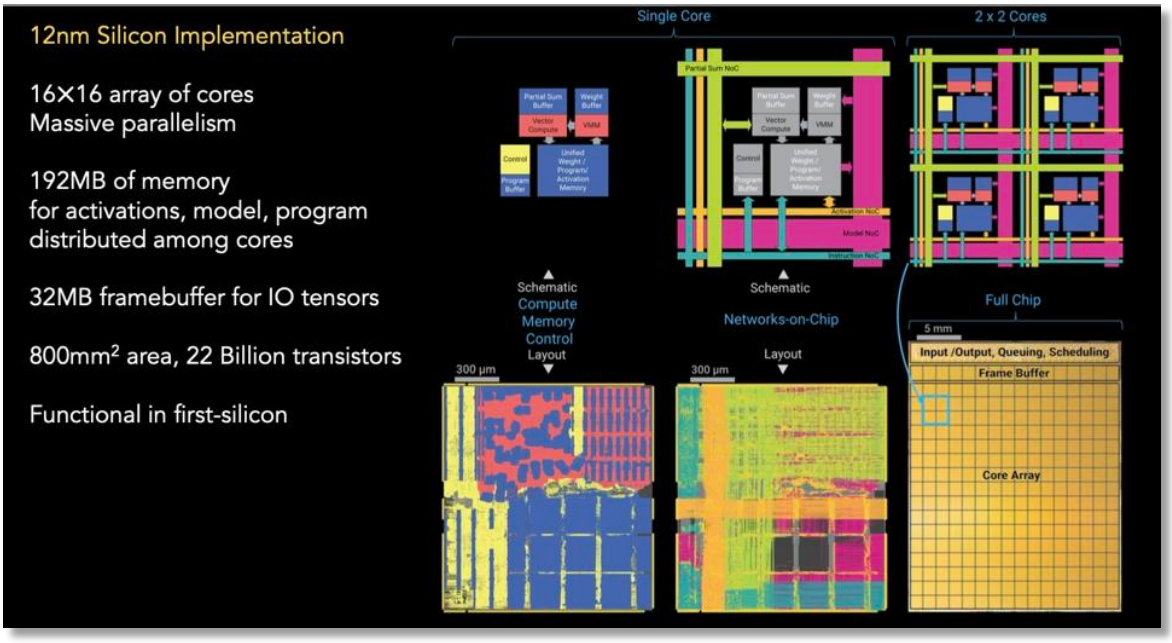
NORTHPOLE ARCHITECTURE

NorthPole is implemented with a novel architecture that differs from traditional computer chips, allowing it to perform AI tasks more efficiently. Here's how NorthPole works:

The NorthPole architecture is designed to be a highly energy-efficient processor specialized for executing inference-based neural networks. Here are some key aspects of the NorthPole architecture:

- **Computational Units**: NorthPole consists of a large array of computational units, typically organized in a grid-like fashion. Each computational unit is capable of performing calculations necessary for inference tasks. These units are optimized for lower-precision calculations, ranging from two to eight bits.
- **Local Memory**: One notable feature of NorthPole is the integration of local memory within each computational unit. This design choice allows the chip to store the weights and connections of the neural network exactly where they are needed, reducing the energy consumption associated with transferring data between memory and processing units.

- **On-Chip Networking**: The architecture includes extensive on-chip networking capabilities, with at least four distinct networks. Some of these networks are used to carry information from completed calculations to the compute units where they are needed next. Others are used to reconfigure the entire array of compute units, providing neural weights and code necessary for executing different layers of the neural network.
- **Parallel Execution**: NorthPole is designed for massive parallel execution. Its computational units can perform thousands of calculations in parallel, which is essential for efficient inference tasks. The absence of conditional branches in the execution units simplifies the hardware and ensures that the wrong code will be executed whenever speculative branch execution turns out to be incorrect.
- **Energy Efficiency**: The architecture is optimized for energy efficiency in executing inference tasks. By eliminating the need for frequent data transfers between memory and processing units and focusing on lower-precision calculations, NorthPole aims to significantly reduce the power consumption associated with AI inference.
- **Specialization**: It's important to note that NorthPole is not a general-purpose AI processor. Instead, it is specifically designed for inference-focused neural networks, such as image classification and audio transcription. It may not handle very large neural networks that do not fit within its hardware constraints.



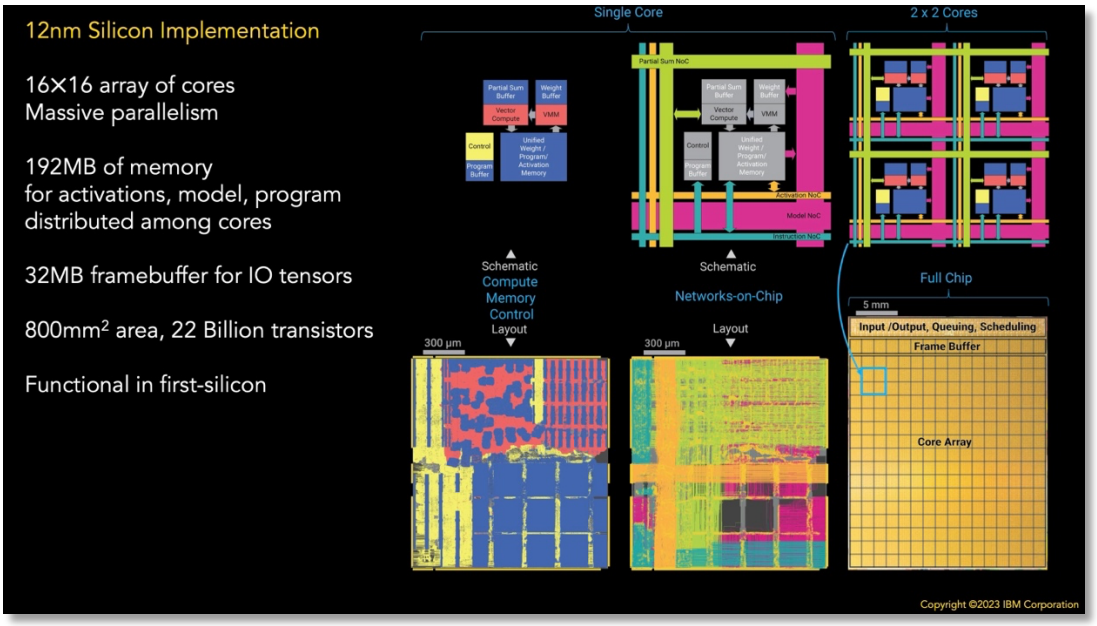
IBM NorthPole Neural Accelerator (IBM CORPORATION)

NorthPole's unique architecture, which integrates processing and memory on the same chip and minimizes data transfer between components, results in higher energy efficiency, lower latency, and improved performance for AI inference tasks.

This chip is designed to be efficient, easy to integrate into systems, and suitable for a wide range of AI applications.

NORTHPOLE TEST CHIP

IBM built multiple test chips to demonstrate the viability of the architecture. The test chips, built on a 12nm process, demonstrated impressive results compared to a Nvidia V100 Tensor Core GPU, performing 25 times the calculations for the same power and outperforming a cutting-edge GPU by about fivefold.



IBM NorthPole Test Chips (IBM CORPORATION)

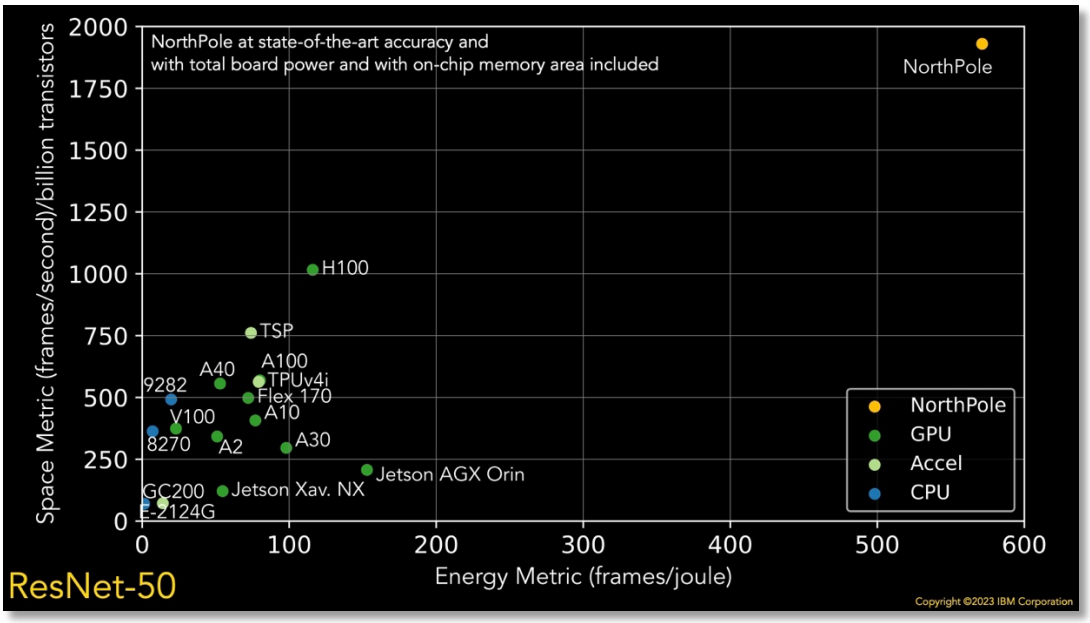
IBM build the test chips with a 16x16 array of cores to provide the level of parallelism required for inference acceleration. It also includes 192MB of memory, and 23MB of framebuffer for I/O tensors.

The resulting chip contains 22 billion transistors over an 800mm^2 area.

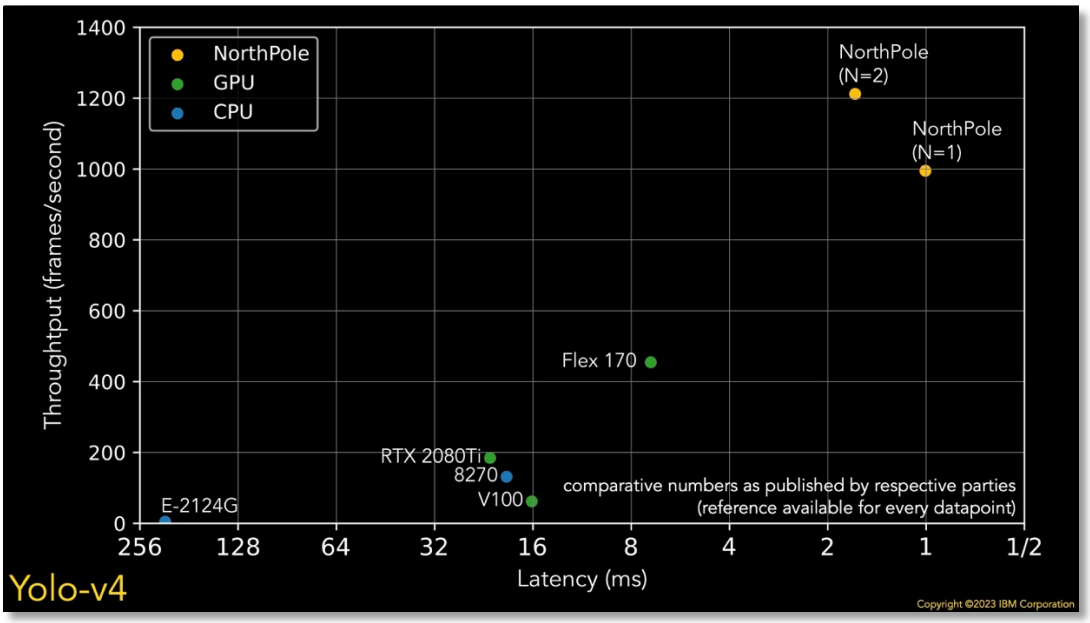
NORTHPOLE PERFORMANCE

IBM's NorthPole demonstrated exceptional performance in tasks like image recognition and object detection, outperforming existing chips in both performance and efficiency.

In tests with AI systems such as ResNet 50 and Yolo-v4, IBM demonstrated that NorthPole is 25 times more energy-efficient and 22 times faster than Nvidia's V100 GPU. Even compared to more advanced nodes like Nvidia's H100 GPU, NorthPole is five times more energy efficient.



NorthPole Energy Comparison: ResNet-50 (IBM CORPORATION)



NorthPole Latency Comparison: Yolo-v4 (IBM CORPORATION)

NorthPole's memory is all on the chip, enabling efficient memory access for each core. This architecture also allows NorthPole to appear as an active memory chip from the outside, simplifying integration into new systems.

NorthPole is optimized for low-precision operations (2-bit, 4-bit, and 8-bit), achieving high accuracy on neural networks while avoiding the high precision required for

training. It operates at a frequency range of 25 to 425 megahertz and can perform 2,048 operations per core per cycle at 8-bit precision.

A standout feature of NorthPole is its ability to process data efficiently without the need for bulky liquid cooling systems, making it suitable for deployment in compact spaces. Ongoing research efforts aim to explore further innovations and advancements in chip processing technologies, promising even greater efficiency and performance gains.

NorthPole has some limitations, such as its inability to handle extremely large models like GPT-4 and its focus on inference rather than training. Nonetheless, the effort represents a significant advancement in AI hardware architecture, offering improved energy efficiency, speed, and memory integration.

ANALYSIS

NorthPole is the culmination of nearly two decades of AI research at IBM Research, focused on creating digital brain-inspired chips. It represents a fusion of traditional processing devices with brain-like processing structures, where memory and processing are intricately intertwined.

The project remained shrouded in secrecy until recently, and its success reflects the dedication and collaborative efforts of the research team at IBM Research. NorthPole signifies a significant milestone in the quest for energy-efficient computing inspired by the human brain.

NorthPole's versatility, high energy efficiency, and ability to handle low-precision operations make it well-suited for various AI applications, including image analysis, speech recognition, and large language models. Its development opens the door to further innovations in AI hardware.

NorthPole is the latest example of IBM's rapid pace of machine learning innovation, which includes solutions such as the Tellum processor in its latest generation z-series and its impressive cadence of Watson.x developments. At the same time, there's no word from IBM on when the technology demonstrated in North will make it into production hardware; rest assured that it's coming.

NAND Research is a technology-focused industry analyst firm providing research, customer content, market and competitive intelligence, and custom deliverables to technology vendors, investors, and end-customer IT organizations.

Contact NAND Research via email at info@nandresearch.com or visit our website at nandresearch.com.