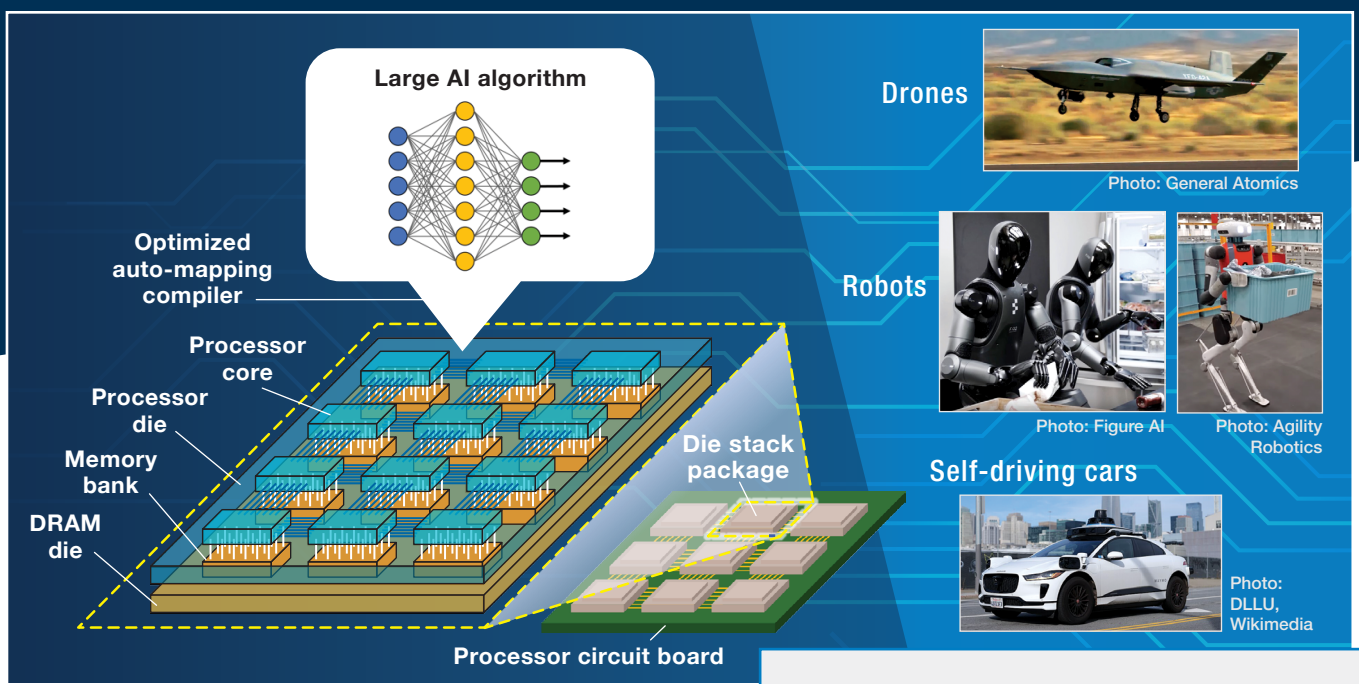
MIT LINCOLN LABORATORY

INNOVATION HIGHLIGHT

Technology in Support of National Security | WWW.LL.MIT.EDU

Empowering Intelligent Autonomous Systems: Novel 3D AI Processor



The 3D AI processor architecture maximizes memory bandwidth with parallel memory access while minimizing the power needed to access memory because of the short interconnects between memory paths.

Researchers at Lincoln Laboratory are developing a 3D processor architecture that can support the huge computational demands of AI employing large language models (LLMs) to make time-critical decisions. Embedded in platforms such as drones, self-driving automobiles, and robots, this compact processor can give these autonomous systems an AI capability similar to a human's ability to quickly assess and respond to a fluid and unforeseen situation.

KEY FEATURES

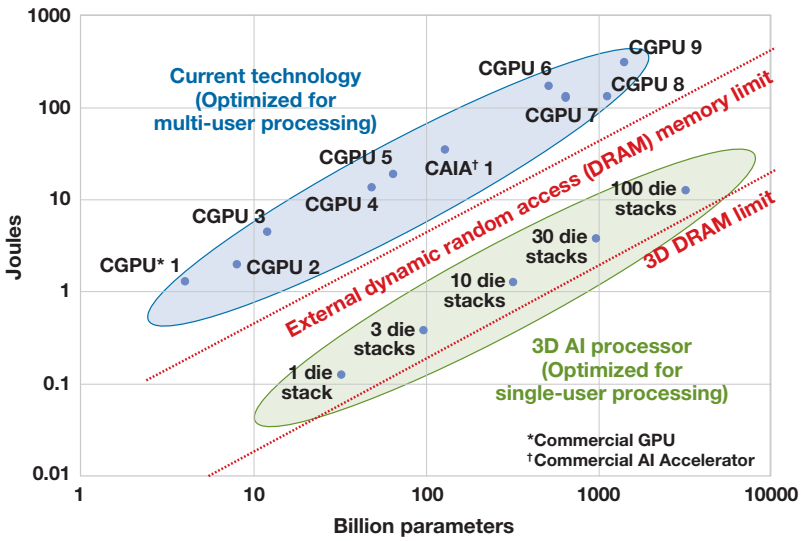
- Low size, weight, power usage, and cost make the processor suitable for integration into autonomous platforms
- Dramatically increased throughput for single-user inferences (i.e., one platform's responses to occurrences) and decreased power consumption are enabled by locating the memory banks beneath a large number of processor cores to provide fast access to memory and to enable parallel computations
- Systolic architecture (synchronized data flow through processor arrays) supports a wide range of neural network models, enabling one hardware platform to handle many AI applications

LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Empowering Intelligent Autonomous Systems: Novel 3D AI Processor

Motivation

Rocks from a landslide are crashing down a hill toward the road ahead of a car. A caregiver is trying to catch an elderly person who has tripped. A military drone needs to determine if there are any noncombatants near a target. A human car driver, caretaker, or drone operator could react immediately to these unexpected changes in the environment. Could the AI application controlling a self-driving car, humanoid assistant, or drone react in time?



Illustrating the 3D AI processor's power efficiency, the graph shows a comparison of the energy per token (inference) vs number of parameters for the 3D AI processor and current commercial GPUs for a single-thread inference.

The increasing use of AI to direct both military and civilian autonomous systems has put a great strain on the processors that must handle the AI computations, especially in time-critical situations. For AI to respond rapidly to untrained situations, developers can use large language models (LLMs), i.e., models trained on vast datasets of text, images, video, and audio, to make human-like inferences, or deductions. To generate inferences from models and their subsequent employment requires data processors with enormous computational throughput, memory capacity, and memory bandwidth. Currently, graphics processing units (GPUs) and other processors with these capabilities are large, costly, and power hungry—qualities that make them unsuitable for embedding within mobile autonomous systems.

Innovative Solution

Technology under development at Lincoln Laboratory is designed to significantly improve throughput while reducing size, weight, power consumption, and costs of a 3D processor for AI used to perform single-user inferences. The concept entails putting the processor die on top of the dynamic random-access memory (DRAM) die so that processor cores optimized for single-user inferences are situated right on top of thousands of equally sized DRAM memory banks. Very short 3D vertical wires provide high-speed, low-power memory access. The cores and die stacks are arranged in 2D systolic array configurations that allow AI developers to implement relatively small LLMs or very large LLMs by utilizing various numbers of identical die stacks.

The embedded 3D AI processor can remain fully operational in environments where wireless communication is unreliable or jammed. This compact processor, which is projected to yield a thousand times more single-user inferences per second per watt while achieving more than a tenfold decrease in manufacturing costs over current GPUs, could enable truly autonomous and immediately responsive mobile systems.

RELATED PATENT APPLICATIONS

PCT/US2024/019512
3D PROCESSOR

PCT/US2024/034736
SYSTOLIC AI PROCESSOR COMPILER

INTERESTED IN WORKING WITH MIT LINCOLN LABORATORY?

Scan the QR code to learn more
www.ll.mit.edu/partner-us

Contact the Technology Transfer Office
tto@ll.mit.edu

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.This material is based upon work supported by the Under Secretary of War for Research and Engineering under Air Force Contract No. FA8702-15-D-0001 or FA8702-25-D-B002. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of War for Research and Engineering.