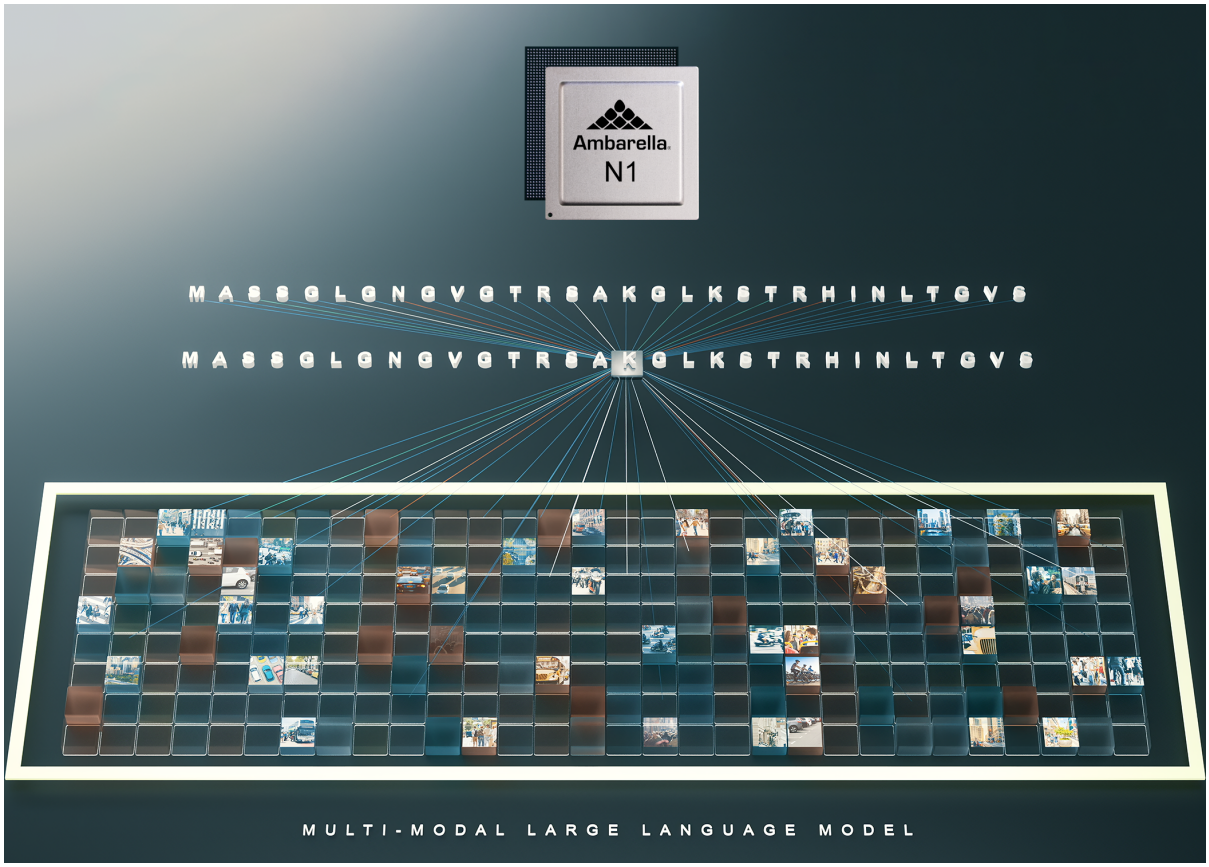
Ambarella Brings Generative AI Capabilities to Edge Devices; Introduces N1 System-on-Chip Series for On-Premise Applications

January 8, 2024 at 5:00 AM EST

Single SoC Supports One to 34 Billion-Parameter, Multi-Modal LLMs With Low Power Consumption, Enabling Generative AI for Edge Endpoint Devices

SANTA CLARA, Calif., Jan. 08, 2024 (GLOBE NEWSWIRE) -- Ambarella, Inc. (NASDAQ: AMBA), an edge AI semiconductor company, today announced during CES that it is demonstrating multi-modal large language models (LLMs) running on its new N1 SoC series at a fraction of the power-per-inference of leading GPU solutions. Ambarella aims to bring generative AI—a transformative technology that first appeared in servers due to the large processing power required—to edge endpoint devices and on-premise hardware, across a wide range of applications such as video security analysis, robotics and a multitude of industrial applications.



Ambarella will initially be offering optimized generative AI processing capabilities on its mid to high-end SoCs, from the existing CV72 for on-device performance under 5W, through to the new N1 series for server-grade performance under 50W. Compared to GPUs and other AI accelerators, Ambarella provides complete SoC solutions that are up to 3x more power-efficient per generated token, while enabling immediate and cost-effective deployment in products.

“Generative AI networks are enabling new functions across our target application markets that were just not possible before,” said Les Kohn, CTO and co-founder of Ambarella. “All edge devices are about to get a lot smarter, with our N1 series of SoCs enabling world-class multi-modal LLM processing in a very attractive power/price envelope.”

“Virtually every edge application will get enhanced by generative AI in the next 18 months,” said Alexander Harrowell, Principal Analyst, Advanced Computing at Omdia. “When moving genAI workloads to the edge, the game becomes all about performance per watt and integration with the rest of the edge ecosystem, not just raw throughput.”

All of Ambarella’s AI SoCs are supported by the company’s new Cooper™ Developer Platform. Additionally, in order to reduce customers’ time-to-market, Ambarella has pre-ported and optimized popular LLMs, such as Llama-2, as well as the Large Language and Video Assistant (LLava) model

running on N1 for multi-modal vision analysis of up to 32 camera sources. These pre-trained and fine-tuned models will be available for partners to download from the Cooper Model Garden.

For many real-world applications, visual input is a key modality, in addition to language, and Ambarella's SoC architecture is natively well-suited to process video and AI simultaneously at very low power. Providing a full-function SoC enables the highly efficient processing of multi-modal LLMs while still performing all system functions, unlike a standalone AI accelerator.

Generative AI will be a step function for computer vision processing that brings context and scene understanding to a variety of devices, from security installations and autonomous robots to industrial applications. Examples of the on-device LLM and multi-modal processing enabled by this new Ambarella offering include: smart contextual searches of security footage; robots that can be controlled with natural language commands; and different AI helpers that can perform anything from code generation to text and image generation.

Most of these systems rely heavily on both camera and natural language understanding, and will benefit from on-device generative AI processing for speed and privacy, as well as a lower total cost of ownership. The local processing enabled by Ambarella's solutions also perfectly suits application-specific LLMs, which are typically fine-tuned on the edge for each individual scenario; versus the classical server approach of using bigger and more power-hungry LLMs to cater to every use case.

Based on Ambarella's powerful CV3-HD architecture, initially developed for autonomous driving applications, the N1 series of SoCs repurposes all this performance for running multi-modal LLMs in an extremely low power footprint. For example, the N1 SoC runs Llama2-13B with up to 25 output tokens per second in single-streaming mode at under 50W of power. Combined with the ease-of-integration of pre-ported models, this new solution can quickly help OEMs deploy generative AI into any power-sensitive application, from an on-premise AI box to a delivery robot.

Both the N1 SoC and a demonstration of its multi-modal LLM capabilities will be on display this week at the Ambarella exhibition during CES.

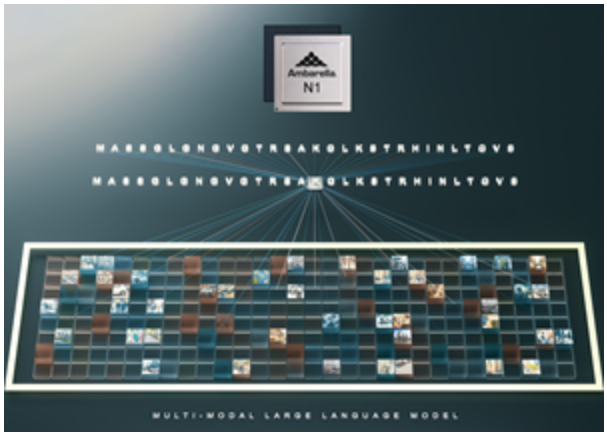
About Ambarella
Ambarella's products are used in a wide variety of human vision and edge AI applications, including video security, advanced driver assistance systems (ADAS), electronic mirror, drive recorder, driver/cabin monitoring, autonomous driving and robotics applications. Ambarella's low-power systems-on-chip (SoCs) offer high-resolution video compression, advanced image and radar processing, and powerful deep neural network processing to enable intelligent perception, fusion and planning. For more information, please visit www.ambarella.com.

Ambarella Contacts

- Media contact: Eric Lawson, elawson@ambarella.com, +1 480-276-9572
- Investor contact: Louis Gerhardy, lgerhardy@ambarella.com, +1 408-636-2310
- Sales contact: https://www.ambarella.com/contact-us/

A photo accompanying this announcement is available at https://www.globenewswire.com/NewsRoom/AttachmentNg/10ee7319-3783-423b-b8f6-d3f6d219293b

Ambarella's New N1 SoC Supports Up to 34 Billion-Parameter, Multi-Modal Large Language Models With Low Power Consumption, Enabling Generative AI for Edge Endpoint Devices



Ambarella is demonstrating multi-modal LLMs running on its new N1 SoC series at a fraction of the power per-inference of leading GPU solutions. The company aims to bring generative AI—a transformative technology that first appeared in servers due to the large processing power required—to edge endpoint devices and on-premise hardware, across a wide range of applications such as video security analysis, robotics and a multitude of industrial applications.