NovuMind: An Early Entrant in AI Silicon

A Look Into The Company's Strategy and Technology

Executive Overview

NovuMind is a Silicon Valley startup that builds full-stack Artificial Intelligence (AI) solutions, AI models, algorithms, boards, chips, and intellectual property for both cloud and edge applications. The company focuses on AI inference where a trained deep neural network is used to process images, sound, speeches and other types of information. The market for inference processing is forecasted to grow rapidly as AI applications become pervasive in robotics, cloud applications, autonomous vehicles, and smart edge devices. This white paper will explore the company's strategy, technology and ability to differentiate in this fast-moving and soon-to-be crowded marketplace.

The company raised ~$15M in Series A funding in 2016 and currently seeks funding from US-based investors. NovuMind received its first silicon product, the NovuTensor chip, back from the foundry in September 2018, and has been awarded four US patents for its core technology.

NovuMind's approach to AI is a Domain Specific Architecture (DSA) that natively performs the 3D tensor computations which are inherent in many deep neural networks, especially the convolutional neural networks (CNNs) used for image processing. NovuMind's claim to unique advantage is that its DSA runs these computations natively, without having to first unfold 3D tensors into 2D arrays for processing, achieving a high degree of parallel processing with far fewer memory operations. The result is superior performance efficiency at low power cost. While the benefits of this approach are being validated by customers, NovuMind has been able to capture the customer attention and secure design wins that support the company's claims of superior performance/watt in edge inference applications.

NovuMind has targeted the fastest growing markets for AI where image processing performance, power and cost are critical. These markets include consumer products, autonomous vehicles, smart cameras, cloud computing, health care and industrial automation. The company has already attracted customers in a few of these markets, and we will explore their use cases in this research note.

We believe that NovuMind has a compelling, and perhaps unique, value proposition for a large emerging market, and has demonstrated the ability to execute with their first-generation silicon. Ren Wu, the company's CEO, has a clear vision of how to address high volume markets for AI inference. According to Wu, upon securing their initial design wins and the next round of funding, the company will build out its executive staff, add engineering capacity and deliver a compelling product roadmap.

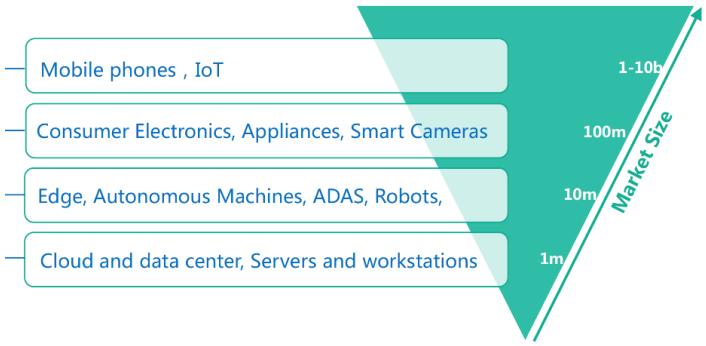
COMPANY OVERVIEW AND STRATEGY

NovuMind was founded in 2015 by CEO Dr. Ren Wu, who had previously been a distinguished scientist at AI pioneer Baidu. Before Baidu, Dr. Wu was the chief software architect at AMD for heterogeneous computing and a principle researcher at Hewlett-Packard, where he became an expert in GPU acceleration of big data analytics on NVIDIA's CUDA. The company now employs about 40 engineers and scientists in their Santa Clara headquarters and in China.

The field of large and small companies vying for attention and leadership in AI acceleration is already quite large, with some 40-50 venture-backed startups in the US, China and the UK. A few of these startups already boast valuations in excess of a billion dollars, but most are still in stealth mode, yet to release their first product. Therefore, to some extent, NovuMind stands out as an early leader simply because they are one of the first to market with working silicon. Their first product, which we shall cover in more detail in the following section, is built on the 28nm manufacturing process at Global Foundries, and the company has shared a technology roadmap to 16 and 7nm process nodes to deliver more performance at lower power and lower costs.

The company's strategy is to provide a portfolio of chips, boards and IP to cover a broad range of markets demanding low power and high-performance image processing. By building a series of chips ranging from ½ watt to 60 watts, the company intends to target multiple markets with a scalable platform. Figure 1 shows the major markets that are in the company's R&D and Go-To-Market plans.

FIGURE 1: NOVUMIND'S VIEW OF THE AI MARKET



NovuMind envisions a family of products that span a variety of form factors, including selling IP to providers of chips for mobile phones, selling chips to integrators of electronics and selling PCIe boards that are used in internet data centers.

Source: NovuMind.

THE 1ST GENERATION OF NOVUTENSOR

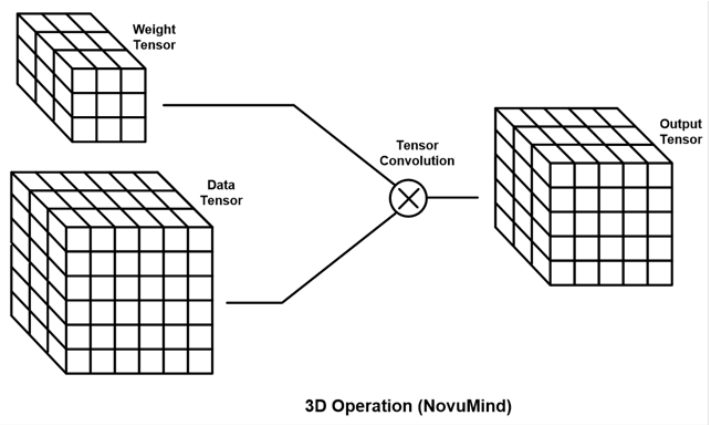
The company is now sampling PCI boards with the 28-nm NovuTensor chip around the world to build a sales opportunity pipeline for this part and future 16- and 7-nm products. The first NovuTensor product is a single-core accelerator that processes convolutional neural networks, the popular AI deep neural network used primarily for image processing. NovuMind claims its first chip delivers 15 Trillion operations per second while consuming 15 watts. The neural processor itself only consumes 5 watts, which is important to note since the company plans to market the logic IP to other chip developers. NovuMind believes that simpler is better, so their chip natively processes (height, width, depth) 3d tensor of an input map, convolved by a set of 3x3xd filters, or 4d (3, 3, depth, N) tensors. This design is far more efficient for the specific operations needed for CNNs, the company contends.

Frankly, direct comparisons are difficult because GPU-based inference parts, such as the NVIDIA T4 GPU and the Xavier autonomous driving System on a Chip (SOC), are far more complex and provide more functions than the NovuTensor part. Those additional functions, of course, take up die area, cost more and consume more power. If someone wants to deploy NovuTensor in an autonomous vehicle, for example, additional devices would be required to provide needed functions such as route planning, sensor fusion, etc.

As a proof point of the advantage of the NovuTensor approach, NovuMind can demonstrate a You Only Look Once (YOLO)-2 object detection model running on

NovuTensor and processing 220 frames per second of video. This means that a single NovuTensor could, in theory, run object detection on seven concurrent 30 fps video camera streams. While we look forward to seeing more application performance data, the initial product seems to have a uniquely strong combination of performance, low power and low cost.

FIGURE 2: NOVUMIND'S 3D OPERATION



NovuMind's architecture uses tensors as its native data type and operations are performed on these 3D matrices.

Source: NovuMind

In addition to delivering the first NovuTensor chip, the company has attracted some top AI scientists and built a large training server infrastructure, called NovuStar, using GPUs to train neural networks it can then use to help customers build custom solutions to run on the NovuTensor device. This is a smart strategy. It positions the company as a partner in solution development, not just a merchant silicon vendor for inference processing. NovuMind also builds AI solutions, which it calls NovuBrain, including networks for object classification, facial recognition, video image enhancement and other AI-enabling capabilities as they develop sales opportunities with end-users.

THE COMPANY'S TECHNOLOGY ROADMAP

As one would expect, NovuMind's engineers are now planning new products to be built on advanced semiconductor process nodes. The company plans their next generation chips to begin shipping later this year. These next generation chips include more cores and higher clock rates yet are designed to maintain the energy and computational

efficiency of its DSA approach. More information can be reviewed under NDA with the company. The company's product plans support the breadth of the company's vision, from embedded AI processing to internet-scale inference processing in the data centers. Common across the product plans is a focus on low cost, low power and high-performance inference applications, which the company believes will become the volume sweet spot for inference ASICs.

Importantly for potential investors, NovuMind's approach keeps the company operating below NVIDIA's radar screens, since they will rarely directly compete. NVIDIA focuses on high value and high performance while delivering at a price and power premium and maintaining utility in a wide variety of programmable workloads. NovuMind focuses on low cost, low power and high performance for a specific set of AI models – a DSA approach.

NovuMind Customer Examples

NovuMind is able to share a few customers' plans for their products, although most customer names cannot be shared without having an NDA in place. Let's look at four examples that demonstrate the company's potential to gain traction in vision-based application domains that demand low cost and low power, yet high performance.

The first customer, Singapore's NCS Pte. Ltd., a subsidiary of telco giant SingTel, provides surveillance systems to enable smart cities. In this example, NovuMind's chip and board products enable advanced video analytics capabilities desired by law enforcement and transit operators. The NovuTensor chip is used in the NCS Claritas family of AI cameras, essentially giving cameras the ability to detect things of interest such as particular vehicles, people or activities. This transforms cameras from mere video capture devices to intelligent endpoints that provide alerts so that authorities can respond to situations immediately. In this application, low power, low cost and low latency performance are critical factors. The performance capabilities of the NovuTensor chip should allow many AI models; e.g. models to detect faces, read license plates, detect suspicious behavior or detect dangerous objects to be processed simultaneously on a given video stream, and all these has to be done within 3w power budget, since the camera is entirely powered by ethernet (PoE). "Our application environment demands high performance at very low power levels," said Kar Han Tan, Head of R&D at NCS. "NovuMind is able to deliver a solution that meets our customers' needs, and at a competitive price-point, thanks to their unique industry-leading architecture."

The second customer use-case example is a large electronics manufacturer that designs smart camera and edge AI systems in the retail and education industries. Once deployed, these systems can help reduce inventory losses, analyze shopping traffic patterns to optimize retail display organization, and monitor student behavior and attention in a classroom environment. The need for low cost and accuracy in processing Convolutional Neural Networks in edge servers is common to these apps. Note that in these applications, being able to process the video in an on-site edge server is also a requirement, due to privacy considerations.

The third example opportunity demonstrates NovuMind's ability to engage with projects that require extremely high-volume, low-power consumer devices. A global leader in high-definition TVs is evaluating NovuTensor's suitability for 8K Super-resolution TV processors. Using AI on NovuMind's chips, lower resolution (4K) images can be up-scaled to 8K images, intelligently filling in the details to produce ultra-quality images while reducing the bandwidth requirements of native 8K image transmissions. In developing this opportunity, NovuMind was the only provider able to pass the customer's stringent requirements for low power and high performance during the initial phase of development and testing. Needless to say, the opportunity here is huge if the company can continue to meet the vendor's needs and secure the funding required to implement the company's technology roadmap.

FIGURE 3: UPSCALING LOWER RESOLUTION IMAGES TO 8K

NovuMind's technology is being tested for suitability in 8K Televisions to up-scale lower resolution images while preserving desired picture quality. This approach has the potential to dramatically reduce the bandwidth required to transmit 8K TV content.

Source: NovuMind.

The final example NovuMind shares is the use of NovuMind in video content analysis in internet-scale data centers (IDCs). Before posting a video on the web, the internet service provider must ascertain the suitability of that content for publication. This is a computationally-demanding application, requiring very high performance to determine whether the video contains objectionable or even illegal content. NovuMind believes their technology can deliver a solution that costs 90% less to deploy and can be operated for 70% less due to its power efficiency. This functionality could eventually be deployed in enterprises as well as the IDC's or provided as a cloud-based application service.

CONCLUSIONS

Unlike many AI silicon startups, NovuMind is not going after the heart of NVIDIA's current or even future business. Rather, the company focuses on the combination of low power, low cost and high performance enabled by their domain specific architecture specifically and exclusively for (inference) processing of Convolutional Neural Networks. Since CNNs for image processing enables a wide range of high-growth applications, this seems like a sound strategy to us. The company's first silicon, built on affordable 28nm fabrication technology, is in production today and has already garnered high-volume design validations and wins. If the company continues to execute well, secures additional funding for their 2nd and 3rd generation chips, executes its product roadmap and builds its sales organization and opportunity pipeline, NovuMind stands a good chance of success.

However, like all startups, the company's potential rewards depend on successfully navigating material risks that must be considered by potential investors or technology adopters. The most significant risk may be the potential of commoditization of inference processing of CNNs. NVIDIA, for example, has open-sourced an ASIC block and IP, called the NVIDIA Deep Learning Accelerator (DLA), for CNN inference processing. This IP can freely be adopted by SOC developers and has been incorporated into NVIDIA's Drive platform. It would appear that NovuMind has a significant performance and energy efficiency advantage over DLA. That fact, combined with NovuMind's early traction with clients, leads us to believe that this risk of commoditization is relatively low, and that NovuMind has a significant opportunity to lead in this fast-growing market, where efficiency and high performance are the key figures of merit.

IMPORTANT INFORMATION ABOUT THIS PAPER

CONTRIBUTOR

Karl Freund, Senior Analyst at Moor Insights & Strategy

PUBLISHER

Patrick Moorhead, Founder, President, & Principal Analyst at Moor Insights & Strategy

INQUIRIES

Contact us if you would like to discuss this report, and Moor Insights & Strategy will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in-context, displaying author's name, author's title, and "Moor Insights & Strategy". Non-press and non-analysts must receive prior written permission by Moor Insights & Strategy for any citations.

LICENSING

DISCLOSURES

This paper was commissioned by NovuMind, Inc. Moor Insights & Strategy provides research, analysis, advising, and consulting to many high-tech companies mentioned in this paper. No employees at the firm hold any equity positions with any companies cited in this document.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Moor Insights & Strategy disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Moor Insights & Strategy and should not be construed as statements of fact. The opinions expressed herein are subject to change without notice.

Moor Insights & Strategy provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.