MICROSOFT'S COBALT 100 PROCESSOR & MAIA 100 AI ACCELERATOR

STEVE MCDOWELL, CHIEF ANALYST
DECEMBER 7, 2023

CONTEXT

Microsoft is undergoing what is described as the largest infrastructure buildout in history, with an annual spend of over $50 billion planned for 2024 and beyond on datacenters. This investment aims to accelerate the path to AGI (Artificial General Intelligence) and integrate generative AI into various aspects of life.

At its recent Microsoft Ignite event, the company announced the launch of the Azure Cobalt 100 CPU and the Maia 100 AI accelerator, marking a significant pivot in Microsoft's approach to its cloud infrastructure and representing the company's commitment to driving innovation in high-performance computing and AI.

This Research Note delves into the intricacies of Microsoft's latest advancements, exploring their potential impact on the future of cloud computing and AI, and looks at how these developments position Microsoft in the ever-evolving public cloud market.

COBALT 100 ARM-BASED PROCESSOR

The Azure Cobalt 100 CPU is a significant development in Microsoft's cloud computing technology, specifically designed for deployment in its Azure cloud services. Here are some key features and aspects of the Cobalt 100 processor:

- **ARM-Based Architecture:** The Cobalt 100 is Microsoft's second ARM-based CPU utilized in its cloud infrastructure, following a previous deployment of an ARM-based CPU from Ampere Computing. This shift to ARM architecture is notable as it represents a move away from traditional x86 architectures predominantly used in server environments.

- **Neoverse N2 Cores:** The processor features 128 Neoverse N2 cores based on the ARMv9 architecture. The Neoverse N2 is known for delivering higher performance compared to its predecessor, the Neoverse N1, with the Cobalt 100 offering approximately 40% better performance.

- **DDR5 Memory Support:** The CPU includes support for 12 channels of DDR5 memory, which is the latest standard in memory technology, offering higher bandwidth and efficiency compared to DDR4.
- **Based on ARM's Genesis CSS:** Cobalt 100 is primarily based on ARM's Neoverse Genesis Compute Subsystem (CSS) Platform. This platform represents a divergence from ARM's traditional business model of only licensing IP. It provides vendors with verified and laid out design components, significantly simplifying the development process of ARM-based CPUs.
- **Dual Genesis Compute Subsystems Integration:** Microsoft has integrated two Genesis compute subsystems into a single CPU for the Cobalt 100. This approach is similar to what has been seen in other ARM-based CPUs like Alibaba's Yitan 710.
- **Rapid Development Cycle:** ARM has indicated the ability to move from project kickoff to working silicon in as little as 13 months, a timeline that is likely applicable to the development of the Cobalt 100, given Microsoft's rapid deployment of this technology.
- **Application in Microsoft's Cloud Infrastructure:** The Cobalt 100 CPU is already being used for various internal Microsoft products, such as Azure SQL servers and Microsoft Teams, indicating its central role in Microsoft's cloud computing services.

The introduction of the Cobalt 100 CPU is part of Microsoft's broader strategy to enhance its cloud infrastructure, particularly focusing on high-performance computing and AI workloads. The new processor represents a key advancement in Microsoft's cloud computing capabilities, leveraging the latest ARM technology to provide enhanced performance and efficiency in its Azure cloud services.

MAIA 100 AI ACCELERATOR

The Azure Maia 100, also known as Athena or M100, is a powerful AI accelerator developed by Microsoft. It represents a major step in Microsoft's efforts to enhance its AI infrastructure within the Azure cloud computing environment. Here are the key features and aspects of the Maia 100:

- **Advanced Manufacturing Technology:** The Maia 100 is fabricated using TSMC's 5nm process technology, which is one of the most advanced semiconductor manufacturing technologies available. This process allows for higher transistor density, resulting in better performance and energy efficiency.

- **High Transistor Count:** The chip boasts a monolithic die with 105 billion transistors, making it one of the largest and most complex chips in terms of transistor count ever publicly disclosed.

- **Impressive Computational Power:** Maia 100 is equipped with significant computational capabilities, offering 1600 TFLOPS of MXInt8 and 3200 TFLOPS of MXFP4. These metrics indicate a strong focus on machine learning and AI performance, particularly for intensive tasks like training large-scale models.

- **Unique Number Formats:** The Maia 100 uses unique number formats - MXInt8 and MXFP4. While these formats are not standard, they are expected to be analogous to existing formats like FP16/BF16 for MXInt8 and FP8 for MXFP4, particularly in inference tasks.

- **Memory Bandwidth:** The chip has a memory bandwidth of 1.6TB/s. While this is impressive, it is less than some of its competitors, which may impact its performance in large language model (LLM) applications that are memory-intensive.

- **Networking Capabilities:** A standout feature of Maia 100 is its built-in RDMA Ethernet IO, allowing for high-speed connectivity directly on the chip. This setup provides a total unidirectional network bandwidth of 4.8Tbps per chip, surpassing many competitors in terms of scale-up bandwidth.

- **Design Considerations for LLMs:** The Maia 100 was designed before the large language model (LLM) craze and therefore has some imbalances, such as a higher proportion of SRAM, which may not be as beneficial for LLM workloads that require substantial off-die memory.

- **Comparative Performance:** In terms of raw FLOPS, Maia 100 competes well with Google's TPUv5 and Amazon's Trainium/Inferentia2 chips, and even with Nvidia's H100 and AMD's MI300X. However, it may face challenges in LLM inference due to its memory bandwidth limitations.

- **Rack and Cluster Architecture:** Microsoft has developed a specialized rack and cluster architecture named Ares for deploying Maia (Athena). This includes custom-designed racks with water cooling systems to accommodate the high power requirements of the Maia chips.

The Azure Maia 100 is a highly advanced AI accelerator that demonstrates Microsoft's commitment to enhancing its capabilities in AI and machine learning. While it shows great promise in terms of computational power and network bandwidth, its performance in certain types of AI workloads, particularly those involving large language models, may be affected by its memory bandwidth constraints.

ANALYSIS

Microsoft's announcements signal a significant strategic advancement in its cloud computing and artificial intelligence capabilities. The Azure Cobalt 100 CPU, evolving from an ARM-based Neoverse N1 CPU and transitioning to a 128-core Neoverse N2 architecture, marks a substantial leap in performance, with a 40% increase over its predecessor.

This shift highlights Microsoft's commitment to leveraging advanced ARM technology for high-performance, efficient cloud computing. The Maia 100 AI accelerator, or Athena, underscores Microsoft's entry into the competitive arena of AI hardware, rivaling products from Google, Amazon, and Nvidia.

With its 5nm process and an impressive 105 billion transistors, the Maia 100 is poised to significantly boost AI and machine learning workloads, despite facing challenges in memory bandwidth that could impact its effectiveness in large language model applications. Additionally, its built-in RDMA Ethernet IO with 4.8Tbps network bandwidth per chip is a notable feature that enhances its scalability and performance in complex AI tasks.

These developments are part of a broader trend where the tier one CSPs increasingly invest in custom silicon to gain a competitive edge in cloud and AI services. Microsoft’s foray into this space, although lagging competitors like Amazon and Google in terms of deployment, is ambitious and could reshape its positioning in the cloud computing market.

Moreover, Microsoft's historical expertise in silicon, evident in its diverse range of previous projects like Project Catapult and custom CPUs for gaming consoles, provides a solid foundation for these latest endeavors.

Microsoft’s focus on internal silicon development, alongside a strategic approach to systems-level design including networking and security, indicates a comprehensive vision for future infrastructure.

The information contained within this document was believed by NAND Research to be reliable and is provided for informational purposes only. The content may contain technical inaccuracies, omissions, or typographical errors. This document reflects the opinions of NAND Research, which is subject to change. NAND Research does not warranty or otherwise guarantee the accuracy of the information contained within.

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