



SOLUTION BRIEF

arm

Bringing Enterprise-Ready AI to Cost-Efficient Compute for Retail



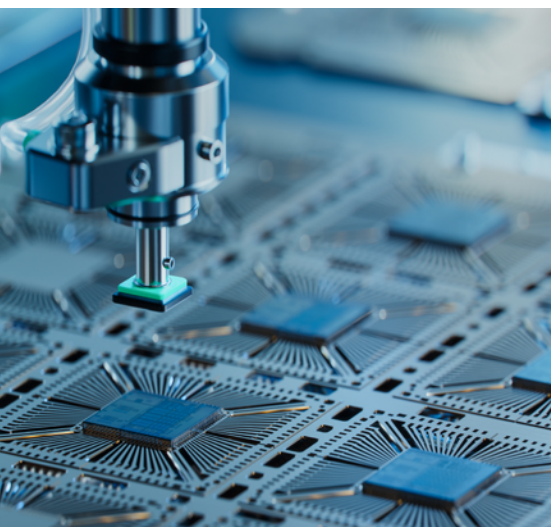
APPLICATION AREA

- 5G
- Artificial Intelligence

OVERVIEW AND GOAL

Arcee AI specializes in small language models (SLMs) optimized for cost-effective inference, making them ideal for enterprise edge applications in Retail, Telco, and Financial Services. They're purpose-built for specific tasks, data, and requirements, with significantly fewer parameters—making them fast, efficient, and cost-effective. They deliver comparable performance to larger models while significantly reducing hardware and operational costs.

To maximize efficiency and scalability, Arcee AI runs their models on Arm-based CPUs and Arm Kleidi - leveraging the unique combination of performance, cost-efficiency, and scalability.



ARM IP

- Cortex-A
- Neoverse
- Kleidi

CHALLENGE:

Large retailers aiming to deploy AI solutions in their stores often encounter significant technical challenges. On the hardware side, integrating AI requires robust, scalable infrastructure capable of handling vast amounts of data in real-time. Existing systems may lack the necessary processing power, storage, and network bandwidth, necessitating costly upgrades. Ensuring seamless integration with legacy systems and maintaining data integrity across platforms is crucial.

Despite these challenges, retailers aim to achieve several outcomes with AI, including enhanced customer experiences through personalized services, optimized inventory management to reduce waste and improve stock availability, and increased operational efficiency by automating routine tasks.

SOLUTION AND BENEFITS:

Arcee AI has conducted benchmarking of its Virtuoso Lite 10-billion parameter model, demonstrating 3-4x acceleration by moving from 16-bit to 4-bit quantization on Arm CPUs while leveraging Arm Kleidi technology.

This delivers significant cost-performance advantages, reducing cloud expenses while maintaining model quality. Rather than relying on expensive and increasingly scarce GPUs, Arcee AI's models run efficiently on Arm-based cloud instances, including those from AWS, Google Cloud, and Microsoft Azure, as well as edge devices and data center hardware.

LINKS TO FURTHER RESOURCES

Arm Success Story and Video:

<https://www.arm.com/company/success-library/made-possible/arcee-ai>

Arcee AI blog:

<https://www.arcee.ai/blog/the-case-for-small-language-model-inference-on-arm-cpus>