The Case for Small Language Model Inference on Arm CPUs

In the dynamic realm of Artificial Intelligence (AI), Small Language Models (SLMs) and Visual Language Models (VLMs) are emerging as indispensable tools for organizations. Their unique blend of performance, cost-effectiveness, and resource efficiency is reshaping the AI landscape. As the demand for AI-driven solutions escalates across industries, SLMs and VLMs present a compelling inference scenario on a variety of hardware, including Arm CPUs. This blog post delves into the advantages and practical applications of running SLM inference on Arm CPUs, underscoring how this approach is set to redefine the reach and cost-effectiveness of AI solutions.

From LLMs to SLMs

The past few years have witnessed a significant shift in the AI landscape, marked by the rise of large language models (LLMs) that have showcased impressive capabilities in natural language understanding and generation. However, the sheer size and computational demands of these models often render them impractical for many real-world applications. This is where the more compact and efficient small language models (SLMs) and visual language models (VLMs) step in, maintaining high levels of accuracy while being more practical. Recent advancements in model architecture and optimization techniques, such as knowledge distillation, have made it possible for Virtuoso Lite, a 10-billion parameter SLM recently released by Arcee AI, to outperform Nova, a 70-billion parameter model also released by Arcee AI in July 2024 and the best open-source model in its size range at the time. This shift toward smaller models is not just about reducing model size; it's about making best-in-class models accessible across a wide range of environments, from edge devices to cloud servers, without the need for expensive AI accelerators.

Privacy, Security, and Compliance

One of the most significant advantages of running SLMs is the enhanced privacy, security, and compliance they offer. Data sovereignty and confidentiality are of the utmost importance in many organizations, especially those in regulated industries such as healthcare, finance, and government. These organizations require complete control over their data and models, often hosting them on-premises or within a private cloud environment to mitigate the risk of data breaches and ensure compliance with data protection regulations.

For example, a telecom company could use an SLM to analyze network data locally, eliminating the need to transmit sensitive information to an out-of-country location, thus strengthening its security and sovereignty posture. This would also ensure that the analysis is performed in real time, enabling faster decision-making and proactive network management. Similarly, an on-site

VLM could be used to analyze video feeds from security cameras, identifying anomalies and potential threats quicker without having to share any data with a remote location.

Tailored Models

Not all AI applications are created equal, and the one-size-fits-all approach to model deployment is becoming less and less viable. SLMs can be tailored to specific tasks and domains, allowing organizations to optimize their models for particular business use cases and to deliver higher returns on investment. This level of customization is crucial to achieve the best possible performance and efficiency, especially when data is highly domain-specific. For instance, in a smart factory, local servers can run tailored SLMs to monitor and predict the maintenance needs of industrial machinery. This unattended operation allows for proactive maintenance scheduling, reducing downtime and extending the life of equipment. Similarly, a local VLM could be used to analyze images from quality control cameras, quickly identifying defects in real time and ensuring that only high-quality products are shipped.

Cost-Performance

Cost efficiency is a key consideration for any technology deployment, and SLMs offer a compelling value proposition. Regardless of the underlying hardware platform they run on, SLMs require fewer computing and memory resources, allowing organizations to reduce their IT spending without compromising prediction quality. Edge devices, in particular, may experience spiky traffic patterns, with periods of intense activity followed by long stretches of idle time. In such environments, the ability to run both the application and the model on the same hardware without the need for a dedicated accelerator can significantly reduce costs.

SLM Inference on CPU

Arm CPUs are everywhere, from smartphones and tablets to servers and cloud instances. This ubiquity means that organizations can leverage existing hardware resources to run SLMs without the need for additional, specialized equipment. The ability to run the model and the application on the same hardware is particularly advantageous in resource-constrained environments, where every watt of power and every byte of memory counts.

For example, the Arm Neoverse platform provides a scalable and efficient architecture for data center and edge computing. Arm designed the Neoverse v1 and v2 processors to deliver high performance and energy efficiency, making them ideal for running SLMs in enterprise environments. These processors support advanced instruction sets that accelerate common deep learning operations, such as matrix multiplication and dot products, further enhancing performance.

To demonstrate the cost-efficiency of Arm CPUs compared to other CPUs, Arcee AI ran inference benchmarks on two comparable Amazon EC2 CPU instances:

- A c8g.8xlarge instance, powered by an AWS Graviton4 CPU (32 vCPUs), itself based on Arm Neoverse V2, and priced at $1.276 an hour (on-demand price in the us-east-1 region).
- A c7i.8xlarge instance, powered by an Intel(R) Xeon(R) Platinum 8488C CPU (32 vCPUS), priced at $1.428 an hour (on-demand price in the us-east-1 region).

We used our Virtuoso-Lite model (10 billion parameters) and quantized it to 4 bits (Q4_0) with the latest source build of the popular llama.cpp open-source project. With 32 vCPUs and at batch size 1, the c8g instance runs inference at 40 tokens per second, while the c7i instance is only able to deliver 10 tokens per second. Factoring in the respective instance costs, this translates into a 4.5x cost-performance advantage for c8g. Of course, your mileage may vary, but this shows how easy it is to get excellent out-of-the-box inference performance with Arm CPUs and open-source tools without the need for vendor add-ons and arcane optimizations.

SLM Inference in Resource-Constrained Environments

For resource-constrained applications running on-device and at the edge, the advantages of SLMs on Arm CPUs become even more pronounced. Spotty connectivity, limited bandwidth, high latency, and data costs can be major hurdles for AI applications. By running models on-site, organizations can overcome these challenges and ensure that their applications are reliable and responsive. For example, utility companies often deploy equipment in isolated areas or even underground, making it difficult or impossible to connect to the cloud. With an on-site SLM, it is possible to leverage edge device data for monitoring, control, and reporting in order to provide maintenance teams with quality and actionable insights. This local processing would ensure that an electrical grid or a water distribution system would respond quickly to anomalies and potential issues, reducing the risk of outages.

SLM Inference in the Cloud

The benefits of SLMs on Arm CPUs also extend to cloud-based deployments. Cloud providers like Amazon Web Services (AWS) and Google Cloud Platform (GCP) offer Arm-based instances, such as the Amazon EC2 Graviton4 and GCP Axion, which deliver high performance at a lower cost. These instances are ideal for running SLMs as they provide the computational power needed for inference while minimizing expenses. Furthermore, the ability to scale resources in or out as required, including scaling to almost zero cost when the application is idle, makes Arm-based cloud instances a cost-effective solution for a wide range of use cases.

For example, in the retail sector, a cloud-based SLM can analyze customer data to provide personalized recommendations and optimize inventory management. By running this model on an Arm-based cloud instance, retailers can achieve high performance at a lower cost, ensuring that their AI applications are both efficient and scalable. Similarly, a VLM could be used to analyze images from in-store cameras, identify customer behavior patterns, and optimize store layouts and product placements.

Many organizations run SLM inference of GPU, and for large-scale, high-throughput applications, this is undoubtedly an excellent and sensible choice. However, some applications may have different requirements, such as smaller scale, spiky behavior, or high sensitivity to cost. For these, cloud-based GPUs are very often over-sized, over-priced, and may even be difficult to procure due to high demand. As SLMs get smaller and smaller, yet more and more accurate, this discrepancy can only grow, making Arm-based inference hard to overlook.

To show this, we ran some more benchmarks on Virtuoso-Lite, our 10-billion parameter model, with an Arm-based and an NVIDIA-based instance on Amazon EC2:

- An r8g.8xlarge instance, powered by an AWS Graviton4 CPU (32 vCPUs), costs $1.885 an hour.
- A g6e.2xlarge instance, powered by an NVIDIA L40S GPU (48GB GPU RAM), and priced at $2.242 an hour.

Running a 4-bit (Q4_0) model with llama.cpp on 32 vCPUs, the r8g instance generates 45 tokens per second. Going down to 8 vCPUs, the instance still produces a very respectable 20 tokens per second, making it possible to run 4 model copies for an aggregated throughput of 80 tokens per second. Meanwhile, with the 16-bit model on vLLM 0.6.6, the g6e instance delivers 35 tokens per second.

This performance extends to larger models, such as Virtuoso-Medium-v2, one of our 32-billion models. As this model is too large to fit on a single NVIDIA L40S, we need to use 2 NVIDIA L40S GPUs running on a g6e.12xlarge instance, which is equipped with a total of 4 GPUs and costs $10.493 an hour. In this configuration, the instance generates 21 tokens per second, allowing two model copies to deliver 42 tokens per second. Meanwhile, with two model copies running on 16 vCPUs each, the r8g instance achieves a total throughput of 21 tokens per second. That's half the performance of the GPU instance but at 18% of the cost!

Of course, the g6e instance would perform better with larger batch sizes and 8-bit quantization. However, these tests show that Arm-based cloud instances are solid contenders for smaller-scale or spiky workloads or cloud regions where GPU availability is limited.

For completeness, we should discuss whether the 4-bit quantization process we applied to our models degraded their original quality. This can be measured by evaluating perplexity, that is to say, a model's ability to accurately predict the next token on a given dataset (lower is better). Our tests show that the larger the model, the larger the perplexity increase. However, for the models we tested in the 8-billion to 32-billion size range, perplexity only increases a few percentage points, which should be unnoticeable in the vast majority of use cases. Amazingly, the 4-bit version of Virtuoso-Lite only induces 1% degradation, making it practically as good as the 16-bit version.

From SLMs to Workflows

The future of AI lies in platforms like Arcee Orchestra, an agentic workflow platform where high-quality SLMs and tools collaborate to perform complex tasks, or Arcee Conductor, a next-generation inference platform using intelligent model routing to query the best model for the task at hand. In both cases, running many SLMs, both off-the-shelf and tailored, in a scalable and cost-efficient manner is essential to achieving the highest levels of accuracy and return on investment. Arm CPUs, with their balance of performance and efficiency, are well-suited to support such workflows. For example, combined with customer information coming from a data store, an SLM could analyze customer chat messages or call transcripts to understand the nature of the issue and route the interaction to the appropriate support agent. Another SLM could predict and diagnose problems based on historical data and current network conditions, suggesting proactive solutions such as firmware updates. A VLM could analyze images or videos sent by customers to verify hardware issues, providing step-by-step visual guidance for troubleshooting, such as resetting a modem. This integrated approach would ensure efficient and scalable customer support, improving resolution times and customer satisfaction.

Conclusion

The convergence of SLMs and VLMs with Arm CPUs marks a significant milestone in the democratization of AI. The benefits of enhanced privacy, security, cost-effectiveness, and flexibility are undeniable. As SLMs and VLMs continue to advance, their deployment on Arm CPUs will unlock new possibilities, making AI not just a theoretical concept but a practical tool accessible to organizations of all sizes across all industries. From telecommunications to retail, manufacturing to healthcare, the synergy of Arm and high-quality SLMs/VLMs will drive innovation, optimize operations, and shape the future of AI-powered solutions.

If you'd like to know more about Arcee AI and how we can help you build best-in-class AI solutions, please visit www.arcee.ai and book a demonstration.