FriendliAI's PeriFlow Accelerates Large Language Model Inference Serving

FriendliAI's PeriFlow Accelerates Large Language Model Inference Serving

REDWOOD CITY, CALIFORNIA, UNITED STATES, November 8, 2023 /EINPresswire.com/ -- FriendliAI, a leading generative AI serving engine company, has released a new version of their inference serving engine PeriFlow. This new release dramatically improves serving performance for large language models (LLMs) and includes features such as quantization support, expanded compatibility, and state caching.

PeriFlow cuts LLM inference serving costs by 40% to 80% compared to existing solutions while providing low latency and high throughput LLM serving. This cutting-edge engine employs various specialized optimizations including FriendliAI's groundbreaking iteration batching, which is protected by patents in the United States and Korea. The FriendliAI team will showcase PeriFlow at the upcoming SC23 and AWS re:Invent 2023 conferences, recognized as leading events for cutting-edge advancements in high-performance computing and cloud computing, respectively.

Alongside a host of new optimizations, this release introduces major new features such as quantization support, multi-adapter support, multi-modal model support, and state caching. Using quantization methods such as Activation-aware Weight Quantization (AWQ) with PeriFlow, users can achieve higher performance running a 70B Llama 2 model with 4-bit weight quantization on just a single NVIDIA A100 80GB GPU than running four GPUs using existing solutions such as vLLM. Users can also run multiple adapters (e.g., LoRAs) fine-tuned for specific use cases on a single GPU. Additionally, PeriFlow can now handle modality inputs other than text, such as images. Lastly, PeriFlow's powerful state caching boasts an additional drastic boost in performance.

Byung-Gon Chun, Founder & CEO of FriendliAI, emphasizes the significance of this milestone for efficient LLM serving, stating "Generative AI is revolutionizing our lives as it enables more creative and productive services. Many organizations are now training or fine-tuning their own models and discovering how costly and painful it is to serve these models at scale for a large user base."

“We need a significant transformation in the way we serve LLMs before organizations will be able to fully harness the potential of their LLMs,” Chun adds. “PeriFlow is the solution. With PeriFlow mitigating the cost concerns of serving these large models, we’re very excited to see what users will build with their LLMs.”

Get Started with PeriFlow Today

With PeriFlow Container, downloadable now from PeriFlow Container Hub, users can try out PeriFlow on their private NVIDIA GPU environment free of charge for four weeks. PeriFlow Cloud beta is also available. Deploy LLMs on PeriFlow in a matter of minutes and automate your LLM serving. Get Started with PeriFlow Today!

About FriendliAI

FriendliAI is a leading provider of cutting-edge inference serving engines for generative AI. Our mission is to enable our customers to serve their generative AI models at high speeds and low costs. For more information, please visit friendli.ai.

Sujin Oh
FriendliAI
press@friendli.ai
Visit us on social media:
LinkedIn
