Demo: WiCi: Wireless Computing Infrastructure

Wei Li*, Yibin Shen*, Zili Meng
Hong Kong University of Science and Technology
Hong Kong

CCS Concepts

• **Networks** → **Local area networks**; • **Information systems** → **Computing platforms**.

Keywords

wireless computing, GPU sharing

ACM Reference Format:
Wei Li*, Yibin Shen*, Zili Meng. 2026. Demo: WiCi: Wireless Computing Infrastructure. In *The 27th International Workshop on Mobile Computing Systems and Applications (HotMobile '26), February 25–26, 2026, Atlanta, GA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3789514.3796248

1 Introduction

The demand for GPU inference is significantly increasing. People are leveraging GPUs to accelerate computation-intensive tasks, including video editing, neural network training, and large language models (LLMs) that require advanced GPU capabilities in daily life. Additionally, the inference stage of LLMs is increasingly dominating GPU usage compared to training, projected to reach 73% by 2028 [1].

While cloud-based GPUs (i.e., accessing LLM services via Web pages) can support LLM inferences, they present several challenges, including unpredictable costs, access difficulties, and privacy concerns. As mentioned earlier, token consumption will continue to grow. To cope with this surging demand, many cloud computing providers are investing heavily in building their cloud computing infrastructure. Anthropic spends 50 — 65% of its revenue on inference expenses [2]. OpenAI has committed $1.15 trillion on cloud infrastructure from 2025 to 2035.

Consequently, there is a movement toward developing edge AI, which performs computations locally on devices such as smartphones and smart appliances. Although Edge-side solutions save costs, they significantly suffer from quality degradation, which prevents the widespread adoption of edge-side inference. The primary reason behind this is the mobility penalty. Due to constraints such as power supply and weight limitations, edge devices have extremely limited VRAM and computing capabilities, resulting in sacrifices in model quality and inference speed. Even if NVIDIA is actively developing mobile GPUs, the gap is estimated to remain around a hundredfold by 2025.

In this paper, we propose a wireless GPU computing infrastructure (WiCi), reducing the inference cost of providers while largely maintaining the response quality and inference speed. We aim to address these challenges by providing wireless access to server-level GPU resources, enabling local inference tasks on edge devices while offloading computations to GPU workstations. This approach seeks to deliver server-GPU-level performance while maintaining low latency, high privacy, and stability.

*Equal contribution. Zili Meng is the corresponding author.

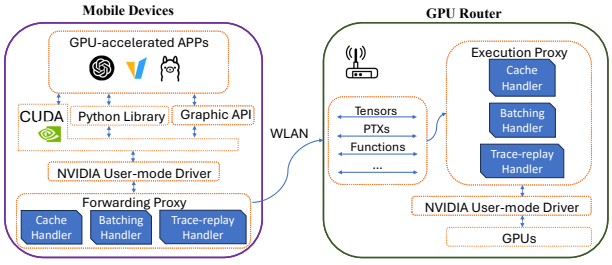

Figure 1: System Overview

2 Design

Fig. 1 illustrates the overview of WiCi. On mobile devices, WiCi intercepts the NVIDIA user driver functions by replacing the runtime library through LD_PRELOAD. It then sends the function calls and input parameters via UDP over Wi-Fi to the server. For each client, the server maintains a dedicated thread to receive packets and execute CUDA functions. When multiple clients are connected, CUDA automatically manages and schedules these requests. Additionally, we can implement our own scheduling algorithm to accommodate customized requirements.

WiCi's implementation includes three primary objectives: an app-agnostic experience, high performance, and scalability. However, challenges include hijacking tasks for remote execution and minimizing networking latency, as Wi-Fi communication is slower than PCIe. To address these challenges, WiCi incorporates designs like model caching, batched execution, WiFi optimizations [3] and trace-replay.

We deploy WiCi on our testbed, which consists of a server equipped with an Intel i7-14700K CPU, 32 GB of memory, and a 4090D 48G GPU. Additionally, we have two client devices, each using a Raspberry Pi 5 with 16 GB of RAM and a 128 GB SD card, both fitted with an Intel AX200NGW network adapter. The server is connected to the router via a wired connection, while the clients connect to the router wirelessly through Wi-Fi. We are now able to run a 40B model on the Raspberry Pi using our system, achieving more than 65% of the native inference speed of the 4090D.

References

[1] China artificial intelligence computing power development assessment report. http://221.179.172.81/images/20250217/25051739782613888.pdf, 2025.
[2] Anthropic - company analysis and outlook report (2026). https://www.deepresearchglobal.com/p/anthropic-company-analysis-outlook-report, 2026.
[3] Yibin Shen and Zili Meng. Law: Towards consistent low latency in 802.11 home networks. In *Proc. USENIX NSDI*, 2026.

HotMobile '26, Atlanta, GA, USA

ACM ISBN 979-8-4007-2471-8/26/02
https://doi.org/10.1145/3789514.3796248