Qi Gao

github/alkalimc

Email: qigaoalkali@sina.com
Mobile: +86-19056824332

Education

- **Southern University of Science and Technology** Shenzhen, China
  Zhiren College *Sept 2024 – Present*

Skills

- **Develop**: Linux, Docker, Mali Driver, frp, SR-IOV, iSCSI, RDMA, PEX
- **LLM**: KTransformers, vLLM, SGLang, llama.cpp, GPTQModel, OpenCompass, lm-eval, EvalPlus
- **Programming**: Shell, Python, JAVA, C/C++, LaTeX
- **Tool**: AutoCAD, KiCad, Altium Designer, RAID/zpool, slurm, ESXi

Experience

- **Metachip Technology Ltd.** Shenzhen, China
  Intern *Feb 2025 – Present*
  - **Server Operations**: Manage the software and hardware management of high-performance deep learning GPU servers in the laboratory. Deployed Slurm for dynamic resource scheduling, optimizing task queuing and execution.
  - **Model Deployment**: Deployed the DeepSeek-R1 model in a resource-constrained environment, optimizing performance for both single- and multi-concurrency tasks. Conducted tuning and optimization to identify the most efficient parameters for deploying a multi-modal model with pipeline and tensor parallelism across multiple frameworks on dual GPUs.

Honors & Awards

- **3rd Prize, The Challenge Cup** SUSTech, Shenzhen, China
  Team Member *Mar 2025*
  - **Project Title**: Design of an Embodied Intelligent Agent Driven by a Large Language Model Integrating Natural Language Understanding and Collaborative Mechanism of Robotic Dog Arm.

Projects

- **FPGA Compiler ToolChain**: Developed a intuitive graphical workflow for model deployment, integrating processes such as loading, quantization, scoring, and deployment. Streamlined the deployment process, reducing complexity and enabling flexible FPGA resource utilization.
- **Multimodal LLM OCR Services**: Optimized the deployment of a 7B multimodal model on dual GPUs under performance constraints by leveraging high-performance INT4 x FP16 Marlin operators and implementing a dual-GPU load balancing workflow.
- **Low-cost Deployment of the DeepSeek R1 model**: Achieved over 10 tokens per second of decoding performance through the use of a CPU-GPU hybrid inference architecture and NUMA-sensitive scheduling, optimizing for cost efficiency.
- **General Evaluation Workflow for Distilled and Base Models**: Developed a comprehensive evaluation workflow for multi-task evaluation comparisons between distilled and base models, enabling quick and thorough analysis of performance differences.
- **Design of an Embodied Intelligent Agent Driven by a Large Language Model Integrating Natural Language Understanding and Collaborative Mechanism of Robotic Dog Arm**: Developed and deployed an adapted vLLM on Nvidia Orin NX, enabling edge-side intelligence for the Unitree Go2 robotic dog. Implemented a multimodal LLM for real-time object recognition and alignment at the end of navigation, utilizing embodied intelligence directly on the edge model.
- **EAIDK 610 Device Tree**: Decompiled the DTBs provided by the Linux mainline, Orangepi, and ARMChina for the 4.x.y and 6.x.y kernel versions, and successfully adapted the system for the 6.15.y kernel. Implemented kernel-space drivers for the Mail GPU and user-space drivers for Panfrost, enabling support for OpenGL 3.2 and Vulkan 1.3 API features. Achieved DirectX 9/11 API compatibility through DXVK translation. Utilized Box64/86 for translating and implementing key Linux amd64 API features, and employed Wine to achieve compatibility with Windows amd64 API features.