

CPE 432: Computer Architecture and Organization (2)

Computer Technology Trends
Research Project

Topic

Investigating AI Accelerator Architectures
and Their Impact on the Performance and
Energy Efficiency of Large Language Model
Inference

Objectives

- Explore **advances in AI accelerator architectures** and their effects on the **performance and energy efficiency of large language model inference**.
- Understand their significance in meeting global computing needs, considering **economic feasibility, environmental impacts, and societal benefits**.
- Examine contemporary design challenges, including **memory bandwidth, low-precision computation, scalability, and cooling**.
- Improve **lifelong learning and critical evaluation skills**.

Procedure

1. Research recent developments in AI accelerators for LLM inference, including **GPUs and specialized accelerators**, focusing on architectural features such as **matrix processing, memory organization, and low-precision arithmetic**.
2. Compare **two accelerator architectures** using published evidence on **inference latency, throughput, and energy efficiency**. Consider the model, numerical precision, batch size, and system configuration when interpreting results.
3. Assess their **economic, environmental, and societal impacts**, including deployment costs, electricity consumption, cooling requirements, and access to AI services.

Plan

1. **Research** the topic from now to the date of Quiz 2 (Tue 29/12/2026)
2. **Take** the assessment (during Quiz 2). Expect two essay questions on the researched technology:
 1. Trends and state-of-the-art developments.
 2. Impacts and effects.

Knowledge Sources

- **Research papers:** IEEE Xplore, ACM Digital Library, and **arXiv**. Note that arXiv preprints may not have undergone peer review.
- **Scholarly search:** Google Scholar and the University of Jordan Library.
- **Technical documentation:** Architecture white papers and developer documentation from accelerator vendors.
- **Benchmark results:** MLCommons **MLPerf Inference**, supplemented by independent evaluations.
- **Supporting resources:** IEEE Computer Society publications, technical courses, and recorded conference talks.
- **AI-assisted research:** Verify claims and references against original sources and disclose AI assistance.

Questions?