Performance Optimization in the LLM World 2024

Kingsum Chow
College of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
kingsum.chow@gmail.com

Yu Tang
College of Software Technology
Zhejiang University
Ningbo, Zhejiang, China
y.tang@zju.edu.cn

Zhiheng Lyu
Department of Computer Science
University of Hong Kong
Hong Kong SAR, China
cogito@connect.hku.hk

Anil Rajput
Datacenter Ecosystem
AMD Corporation
Portland, Oregon, USA
Anil_Rajput@yahoo.com

Khun Ban
Datacenter and AI
Intel Corporation
Hillsboro, Oregon, USA
khunban@gmail.com

ABSTRACT

The popularity and adoption of large language models (LLM) like ChatGPT has evolved rapidly. LLM pre-training is expensive. ChatGPT is estimated to cost over $700,000 per day to operate and using GPT-4 to support customer service can cost a small business over $21,000 a month. The high infrastructure and financial costs, coupled with the specialized talent required, make LLM technology inaccessible to most organizations. For instance, the up-front costs include the emissions generated to manufacture the relevant hardware and the cost to run that hardware during the training procedure, both while the machines are operating at full capacity and while they are not. The best estimate of the dynamic computing cost in the case of GPT-3, the model behind the original ChatGPT, is approximately 1,287,000 kWh, or 552 tons of carbon dioxide. The goal of this workshop is to address the urgency of reducing energy consumption of LLM applications, by bringing together researchers from the academia and industry to share their experience and insights in performance engineering in the LLM world.

ACM Reference format:

Kingsum Chow, Yu Tang, Zhiheng Lyu, Anil Rajput and Khun Ban. 2024. Performance Optimization in the LLM World. In *Companion of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE'24 Companion), May 7–11, 2024, London, United Kingdom.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3629527.3651436

ICPE '24 Companion, May 7–11, 2024, London, United Kingdom

ACM ISBN 979-8-4007-0445-1/24/05.
https://doi.org/10.1145/3629527.3651436

Organiers/presenter and affiliations (including short bios)

Kingsum Chow (kingsum.chow@gmail.com) is a professor at the School of Software Technology, Zhejiang University. He received his Ph.D. in Computer Science and Engineering at the University of Washington in 1996. Prior to joining Zhejiang University in 2023, Kingsum has been working as a chief scientist and senior principal engineer in the industry. He has extensive experience in software hardware co-optimization from thirty years of working at Intel and Alibaba. He delivered two QCon keynotes. He appeared four times in JavaOne keynotes. He has been issued 30 patents. He has delivered more than 100 technical presentations. He has collaborated with many industry groups, including groups at Alibaba, Amazon, AMD, Ampere, Appeal, Arm, BEA, ByteDance, Facebook, Google, IBM, Intel, Microsoft, Netflix, Oracle, Siebel, Sun, Tencent and Twitter. In his spare time, he volunteers to coach multiple robotics teams to bring the joy of learning Science, Technology, Engineering and Mathematics to the K-12 students in USA and China.

Yu Tang (y.tang@zju.edu.cn) is a postgraduate student at the Zhejiang University. His advisor is Kingsum Chow. His research interest focuses on Large Language Model, system performance analysis and optimization.

Zhiheng Lyu (cogito@connect.hku.hk) is a distinguished senior at the University of Hong Kong (HKU) and concurrently serves as a research assistant within the Berkeley AI Research Lab at the University of California, Berkeley. His academic journey is punctuated by seminal contributions at both UCB and ETH Zürich, resulting in key papers on LLM interpretability and practical applications — several of which are under active conference review. A noteworthy internship at Megvii's R-face Institute allowed him to advance automatic CV model training systems, setting new industry standards. Beyond research, Zhiheng's prowess in algorithmic competitions is evident: he boasts two gold medals from regional ICPC events and an impressive appearance in the ICPC

world finals. As the current leader of the HKU AI4Good Community, Zhiheng consistently synthesizes his deep knowledge in AI, robotics, and LLMs to spur innovation. His role as the organizer of this workshop underscores his dedication to fostering richer insights into LLM application and optimization.

Anil Rajput (Anil_Rajput@yahoo.com) is an AMD Fellow, Software System Design, as core architect for datacenter and cloud with focus on performance, deployments, optimizations, and best practices. He received his certification in data analytics from Harvard Business Analytics Program in 2022 and his Master's in Electrical and Computer Engineering from Portland State University in 1997. Currently, Anil's focus areas are workloads characterization, platform evaluation, cloud deployments, on-prem datacenters as well as understanding and resolving large deployment issues at scale for critical customers. Earlier, he has been at Intel Corporation for more than 20 years, playing various roles in the Software and Services Group, leading platform design, managed runtime like Java and .Net, scripting languages and development of representative benchmarks as chair of Java committee at SPEC. He was key members of teams who architected and developed several benchmarks like SPECjbb2005, SPECjvm2008, SPECjEnterprise2010, SPECpower_ssj2008 etc. Anil is also guiding graduate students as mentor and also participates in local High School science fairs to encourage kids for STEM in Oregon, USA.

Khun Ban (khunban@gmail.com) is an Intel cloud performance architect leading a team to drive solutions to solve today's complex business problems by analyzing the requirements and making architecture recommendations for CPUs/storage/network balance to best meet the needs based on the constraints. He has over twenty years of enterprise software development experience. His current focus is on Open-Source Relational Databases. He received his B.S. degree in Computer Science and Engineering from the University of Washington in 1995.

A list of topics to be covered, including format (e.g., talks, demos, etc.), target audience, and pre-requisite knowledge

The half day workshop will be composed of invited talks, work in progress and fully refereed papers and a panel. Presentations are not limited to the following topics:

1. Optimizing LLM Workloads on Traditional and New Architectures
   - Hardware Assisted LLM Systems
   - LLM Optimization at Scale
   - Code generation optimization for modern hardware
2. Panel Discussion (speakers from the industry and academia)

The target audience:

1. Researchers that are advocating new ways of optimizing LLM applications in software or hardware optimizations.
2. Practitioners that need to solve runtime performance problems in their LLM deployments.

Expected duration : half day

Expected attendees: 30

The main organizer delivered the following workshops and tutorials in the past

- Runtimes in the Cloud 3, a full day workshop at HPCA, 2020/02, 20 attendees.
- Runtimes in the Cloud 2, a full day workshop at ISCA, 2019/06, 20 attendees.
- Runtimes in the Cloud, a full day workshop at ISCA, 2018/06, 30 attendees.
- Scaling Software Performance and Software Performance in the Cloud, a half day workshop at PNSQC, 2017/10, 50 attendees.
- Software Performance Analytics in the Cloud, a full day tutorial at ICPE, 2017/04, 20 attendees.
- Applying Analytics to Data Center Performance, a half day workshop at CMG Performance and Capacity Conference, 2015/11, 30 attendees.

Workshop Website

https://sites.google.com/view/pollmw