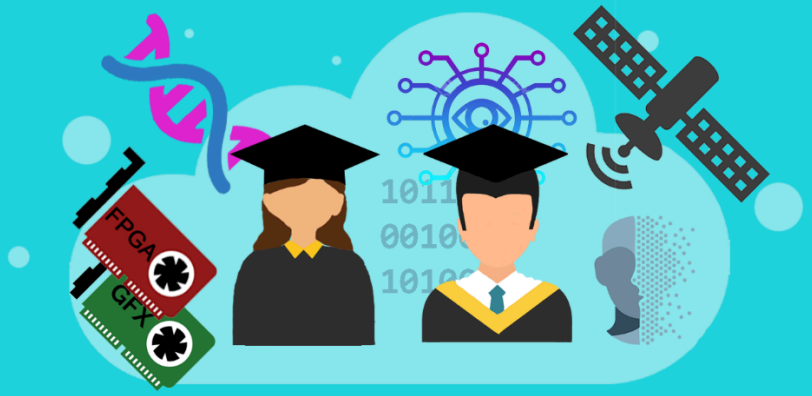




Characterization and Efficient Resource Management of Agentic AI Workloads

Artificial intelligence is increasingly expanding beyond standalone large language model (LLM) interactions toward agentic systems that perform tasks through sequences of reasoning, action, and feedback [1]. Examples include coding agents that modify and test software, research assistants that gather and synthesize information, and automation agents that coordinate workflows across applications. These systems typically operate through an iterative loop: an LLM determines the next action, an external tool executes it, and the resulting output is incorporated into subsequent LLM calls until the task is completed.

Agentic workloads introduce challenges that differ from those of conventional LLM inference. Their execution graphs evolve dynamically, as intermediate results determine subsequent actions, making computational demand and completion time unpredictable. Repeated interactions can also produce long contexts, increasing memory requirements and inference costs [2, 3]. Meanwhile, tool execution introduces waiting periods during which inference resources may remain idle, reducing utilization and potentially wasting energy. Differences in tool latency, dependencies between actions, and opportunities for parallel execution further complicate resource allocation and performance optimization.

This thesis will investigate the unique execution characteristics of agentic AI workloads and explore methods for addressing the challenges they impose on computing systems. Particular attention will be given to the interaction between LLM inference and tool execution, the evolution of context and resource requirements, and the variability of execution paths. The study will explore workload-aware scheduling, memory management, and adaptive power management as potential approaches to improving resource utilization and energy efficiency while maintaining acceptable task completion times.

RELATED MATERIAL:

- [1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," in *Proc. International Conference on Learning Representations (ICLR)*, 2023.
- [2] J. Kim, B. Shin, J. Chung and M. Rhu, "The Cost of Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI Infrastructure Perspective," *2026 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, Sydney, Australia, 2026.
- [3] M. Luo et al., "Agentix: An Efficient Serving Engine for LLM Agents as General Programs," in *Proc. 23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, 2026.

CONTACT INFORMATION:

Christos Kavallaris, Ph.D. Student, NTUA: (ckavallaris@microlab.ntua.gr)
Achilleas Tzenetopoulos, Post Doc, NTUA
Dimosthenis Masouros, Post Doc NTUA
Sotirios Xydis, Ass. Prof., Microlab NTUA
Dimitrios Soudris, Professor Microlab NTUA