

Проблемы и подходы к управлению памятью в высокопроизводительных вычислительных системах

Петросян Амиран Григорович

E-mail: petrosyan.amiran@gmail.com

Современные высокопроизводительные системы и облачные платформы сталкиваются с ограничением memory wall, когда рост мощности процессоров опережает развитие подсистемы памяти. Это снижает эффективность приложений при обработке больших данных, графовой аналитике и ИИ. В традиционных монолитных архитектурах часть памяти остаётся неиспользованной, что препятствует её гибкому перераспределению.

Обзор подходов к управлению памятью включает деагрегацию ресурсов и формирование масштабируемых пулов. Рассмотрены методы прямого доступа к удалённой памяти, аппаратной когерентности и динамического распределения данных, адаптивные стратегии размещения и миграции с учётом нагрузки, а также оптимизация структур данных и вычисления вблизи памяти для повышения локальности [2,3,6,7]. Такие подходы позволяют минимизировать задержки при доступе к данным и повысить пропускную способность.

Особое внимание уделено когерентности, синхронизации и отказоустойчивости, включая восстановление транзакционной целостности, репликацию и продолжение работы при частичных сбоях. Рассмотрены методы моделирования и симуляции для оценки производительности [1,3,5].

Анализ показывает, что динамическое управление памятью повышает масштабируемость и эффективность НРС-систем. Перспективы связаны с разработкой интеллектуальных алгоритмов распределения памяти, кэширования и совместным проектированием аппаратных и программных решений [3,4,5]. Интеграция адаптивных кэш-механизмов и высокопроизводительных ключ-значение хранилищ становится важным направлением для дата-центров нового поколения.

Источники и литература

- 1) Aguilera M.K., Ben-David N., Guerraoui R., Murat A., Xygkis A., и др. uBFT: Microsecond scale BFT using Disaggregated Memory [Extended Version] // arXiv:2210.17174. 2022. DOI: 10.48550/arXiv.2210.17174.
- 2) Lee S., Ponnappalli S., Singhal S., Aguilera M.K., Keeton K., Chidambaram V. DINOMO: An Elastic, Scalable, High Performance Key Value Store for Disaggregated Persistent Memory // Proc. VLDB Endowment. 2022. Vol. [U+202F] 15, No. [U+202F] 13. P. [U+202F] 4023–4037.
- 3) Liu Y., Xie M., Shi S., Xu Y., Litz H., и др. Outback: Fast and Communication efficient Index for Key Value Store on Disaggregated Memory // Proceedings of the VLDB Endowment. 2024. № [U+202F] 2 (18). С. [U+202F] 335–348.
- 4) Murat A., Burgelin C., Xygkis A., Zablatchi I., Aguilera M.K., Guerraoui R. SWARM: Replicating Shared Disaggregated Memory Data in No Time // Proceedings of the ACM SIGOPS 30th Symposium on Operating Systems Principles (SOSP '24). 2024. С. [U+202F] 24–45. DOI: 10.1145/3694715.3695945.
- 5) Shen J., Zuo P., Luo X., Su Y., Gu J., и др. Ditto: An Elastic and Adaptive Memory Disaggregated Caching System // SOSP. 2023. P. [U+202F] 675–691.

- 6) Tang Y., Lee S., Bhattacharjee A., Khandelwal A. PULSE: Accelerating Distributed Pointer-Traversals on Disaggregated Memory (Extended Version) // arXiv:2305.02388. 2023. DOI: 10.48550/arXiv.2305.02388.
- 7) Wang J., Li C., Wang T., Guo J., Yang H., и др. Survey of Disaggregated Memory: Cross layer Technique Insights for Next-Generation Datacenters // arXiv:2503.20275. 2025. DOI: 10.48550/arXiv.2503.20275.