

ТЕМПЕРАТУРНО-АДАПТИВНЕ КЕРУВАННЯ ЗАТРИМКАМИ ПАМ'ЯТІ DDR У СУЧАСНИХ ОБЧИСЛЮВАЛЬНИХ СИСТЕМАХ

Для керування сучасними електронними обчислювальними пристроями та серверними комплексами ключовим критерієм є забезпечення максимальної пропускної здатності підсистеми оперативної пам'яті за мінімально можливих рівнів затримок. Сучасна оперативна динамічна пам'ять стандартів DDR4 та DDR5 функціонує на надвисоких тактових частотах, що неминуче призводить до інтенсивного тепловиділення всередині напівпровідникових чипів під тривалим робочим навантаженням. Підвищення температури кремнієвих кристалів експоненційно прискорює диференційні струми витоку в конденсаторах мікроскопічних комірок пам'яті DRAM. Це явище вимагає значного скорочення інтервалів між циклами примусової регенерації (Refresh) і, як наслідок, призводить до суттєвого падіння загальної продуктивності та пропускної здатності обчислювальної підсистеми. У даній оглядово-дослідній статті детально розглядається проблематика виникнення температурного тротлінгу в сучасних високих профілях пам'яті та пропонується комплексна концепція динамічного температурно-адаптивного керування ключовими затримками (таймінгами) та інтервалами оновлення. Суть запропонованого підходу полягає у розробці та теоретичному обґрунтуванні програмно-апаратного механізму, який у реальному часі на основі телеметрії вбудованих термодатчиків SPD динамічно коригує значення основних (t_{CL} , t_{RCD} , t_{RP} , t_{RAS}) і другорядних таймінгів, а також параметрів t_{REFI} та t_{RFC} . Практичне впровадження такого адаптивного комплексу дозволяє гнучко знижувати затримки на 5–12 % у нормальних температурних діапазонах (до 50°C) і одночасно гарантувати абсолютну відмовостійкість та стабільність системи при екстремальних термічних навантаженнях, оптимізуючи швидкодію комп'ютерних систем без ризику втрати чи спотворення даних.

Ключові слова: оперативна пам'ять, динамічна пам'ять, DDR4, DDR5, затримки пам'яті, таймінги, температурно-адаптивне керування, струми витоку, інтервал оновлення, t_{REFI} , продуктивність обчислювальних систем.

Вступ та постановка проблеми. На сучасному етапі розвитку обчислювальної техніки спостерігається значний дисбаланс між стрімким зростанням швидкодії багатоядерних центральних процесорів (CPU) та відносно повільним прогресом у зниженні латентності оперативної пам'яті (DRAM). Ця проблема, класично відома в архітектурі ЕОМ як "стіна пам'яті" (memory wall), суттєво обмежує реальну продуктивність суперкомп'ютерів, серверів баз даних та високопродуктивних робочих станцій [4, 6]. Для подолання цього бар'єру розробники змушені постійно нарощувати тактові частоти шини пам'яті та щільність компонування напівпровідникових елементів у чипах пам'яті стандартів DDR4 та DDR5. Проте зворотним боком такого екстенсивного підходу є різке збільшення питомого тепловиділення модулів пам'яті [1, 2, 8].

Фізична структура сучасної динамічної пам'яті базується на використанні мікроскопічних конденсаторів та керуючих польових транзисторів. Збереження логічного стану комірки (0 або 1) напряму залежить від наявності або відсутності електричного заряду на відповідному конденсаторі. Внаслідок недосконалості техпроцесів літографії та фізичних властивостей кремнію, цим конденсаторам притаманні постійні струми витоку [2, 7]. Найбільш критичним фактором, що впливає на інтенсивність витоку заряду, є робоча температура напівпровідникового кристала. При її підвищенні струми витоку зростають за експоненційним законом, що призводить до прискореної деградації збереженої інформації.

Згідно зі специфікаціями комітету стандартизації JEDEC, для забезпечення цілісності даних при досягненні температури чипів понад 85°C контролер пам'яті зобов'язаний у жорсткому примусовому порядку скоротити інтервал між циклами регенерації (параметр

tREFI) вдвічі – з базових 7.8 мікросекунд до 3.9 мікросекунд (або відповідно для архітектури DDR5). Частіші цикли оновлення призводять до того, що шина пам'яті тривалий час залишається заблокованою для виконання стандартних операцій читання (Read) та запису (Write) [1, 2, 6]. Це викликає так званий термічний тротлінг підсистеми пам'яті, суттєво збільшуючи середній час відгуку системи.

Існуюча проблема посилюється тим, що параметри затримок (основні таймінги, такі як CAS Latency (tCL), RAS to CAS Delay (tRCD), Row Precharge Time (tRP) та Row Active Time (tRAS)) встановлюються статично під час ініціалізації системи (POST-перевірки BIOS/UEFI) на основі зашитих у мікросхему SPD (Serial Presence Detect) консервативних профілів або екстремальних статичних профілів розгону (XMP/EXPO) [1, 6, 7]. Вони розраховуються на найгірший сценарій експлуатації (worst-case scenario) при максимальній температурі (наприклад, 85°C або 95°C). Таким чином, коли комп'ютерна система працює у штатному режимі з помірними температурами модулів пам'яті (від 35°C до 50°C), затримки залишаються штучно завищеними, що призводить до втрати значного потенціалу продуктивності системи. Отже, розробка гнучких, динамічних програмно-апаратних засобів температурно-адаптивного керування підсистемою пам'яті DDR є вкрай актуальною науково-практичною задачею.

Аналіз останніх досліджень і публікацій. Дослідженню архітектурних особливостей та оптимізації функціонування підсистеми оперативної пам'яті присвячено велику кількість наукових праць. Вагомий внесок у розробку методів підвищення пропускної здатності DRAM та зменшення енерговитрат зробили такі провідні закордонні вчені, як Онур Мутлу (Onur Mutlu), К. Лі (K. Lee), Д. Вонг (D. Wong) та інші [3, 4, 6, 7]. У їхніх роботах детально проаналізовано математичні моделі поведінки комірок пам'яті при масштабуванні техпроцесів та запропоновано різноманітні підходи до оптимізації циклів регенерації, включаючи концепції інтелектуального по-банкового оновлення (Per-Bank Refresh) та мінімізації негативного впливу ефекту RowHammer.

Проте більшість існуючих комерційних та академічних рішень зосереджена на чисто апаратних змінах у структурі самих кристалів DRAM або логіці інтегрованих контролерів пам'яті (IMC) всередині CPU, які потребують тривалого циклу проектування та суттєвих капіталовкладень. Наявні на ринку технології терморегуляції зводяться переважно до дискретного перемикавання станів енергоспоживання або аварійного зниження робочої тактової частоти пам'яті при перегріві. Водночас можливості гнучкого динамічного регулювання конфігурації часових затримок (таймінгів) у реальному часі на основі безперервного моніторингу теплового поля модулів залишаються малодослідженими. Відсутні цілісні математичні моделі та алгоритми, які дозволяли б операційній системі або низькорівневому вбудованому ПЗ (firmware) безпечно взаємодіяти з контролером пам'яті для тонкого підлаштування часових діаграм без ризику виникнення апаратних збоїв (BSoD, Kernel Panic) чи пошкодження структури файлових систем [2, 3, 8].

Формулювання цілей статті. Метою даної статті є комплексний аналіз причинно-наслідкових зв'язків між поточною робочою температурою мікросхем динамічної пам'яті DDR4/DDR5 та їх гранично стабільними швидкісними характеристиками, а також теоретичне обґрунтування і розробка концептуальної математичної моделі програмно-апаратного комплексу для температурно-адаптивного динамічного керування затримками пам'яті в реальному часі. Досягнення поставленої мети передбачає формалізацію аналітичних залежностей основних таймінгів від температурних показників SPD та формування матриці оптимізованих станів для інтеграції у сучасні обчислювальні архітектури [2, 6, 7].

Виклад основного матеріалу дослідження. Фізичні засади функціонування кремнієвих структур DRAM визначають, що час утримання заряду (retention time) окремої комірки зменшується майже вдвічі на кожні 10°C зростання температури кристала вище порогу в 45°C. Це зумовлено термоемісією електронів та зростанням підпорогових струмів витоку через затвор пасивуючого транзистора комірки. Традиційна спрощена схема розрахунку абсолютної латентності операції випадкового доступу до пам'яті базується на комбінації тактової частоти шини та безрозмірних коефіцієнтів затримок (виражених у кількостях тактів). Проте фізичний

час проходження сигналів у шинах та час відкриття внутрішніх сторінок (Row Buffers) змінюється під впливом температури, оскільки рухливість носіїв заряду в кремнії падає при нагріванні [1, 5, 8].

У ході проведеного дослідження було розроблено концепцію замкненого контуру керування (closed-loop control system), яка інтегрується безпосередньо у логіку інтегрованого контролера пам'яті процесора. Система використовує високоточні цифрові термодатчики, інтегровані в мікросхеми SPD Hub (стандарт DDR5) або розпаяні безпосередньо на друкованій платі модулів пам'яті (стандарт DDR4). Періодичність опитування датчиків становить $N = 100\text{ } \mu\text{s}$, що є цілком достатнім з огляду на теплову інерційність масивних радіаторів охолодження модулів оперативної пам'яті [3, 4, 6].

На основі зібраних аналітичних даних було сформовано систему рівнянь та умов адаптації, які згруповані та представлені у вигляді Таблиці 1. Даний математичний апарат покладено в основу алгоритму роботи адаптивного контролера.

Таблиця 1 – Математичні моделі та критерії температурно-адаптивного керування параметрами DRAM

Керований параметр підсистеми	Аналітична формула для розрахунку (Equation)	Фізична умова активації та межі адаптації
Абсолютна латентність (L), нс	$L = (t_{CL} \times 2000) / \text{DataRate}$	Використовується для розрахунку базової часової затримки при еталонній температурі $T = 40^\circ \text{C}$.
Інтервал оновлення (tREFI), мкс	$t_{REFI} = 7.8$, при $T \leq 85^\circ \text{C}$ $t_{REFI} = 3.9$, при $T > 85^\circ \text{C}$	Динамічне ступеневе ущільнення циклів оновлення для запобігання спотворенню зарядів у комірках.
Адаптивна затримка CAS (tCL_adj), тактів	$t_{CL_adj} = t_{CL_base} + \lfloor k_t \times (T - 45) \rfloor$	Активується при $T > 45^\circ \text{C}$; k_t – термічний коефіцієнт деградації сигналу (для DDR4 $k_t \approx 0.045$).
Час попереднього заряду (tRP_adj)	$t_{RP_adj} = t_{RP_base} - \lfloor \alpha \times (45 - T) \rfloor$	Оптимізаційне зниження затримки у "холодній" зоні при $T < 45^\circ \text{C}$ (коефіцієнт прискорення $\alpha = 0.06$).

Як детально демонструють наведені у Таблиці 1 математичні залежності, система забезпечує дворівневий характер регулювання [1, 2]. У зоні низьких та помірних температур (до $45\text{--}50^\circ \text{C}$), коли витоки струму мінімальні, а провідність каналів польових транзисторів є максимальною, алгоритм здійснює агресивне зниження значень tCL та tRP нижче номінального JEDEC-рівня. Це дозволяє виграти дорогоцінні наносекунди при випадковому доступі до пам'яті. У випадку ж інтенсивних обчислень (наприклад, рендеринг відео, навчання нейромереж або робота складних ігрових рушіїв), коли температура починає стрімко наближатися до критичних $80\text{--}85^\circ \text{C}$, система плавно збільшує значення таймінгів на декілька тактів за допомогою оператора округлення до більшого цілого [3, 5, 7]. Це дещо знижує пікову

пропускну здатність, але повністю нівелює ризик виникнення апаратних збоїв чи пошкодження даних.

Для підтвердження ефективності запропонованого методу було проведено серію імітаційних моделювань роботи підсистеми оперативної пам'яті DDR4-3200 (двоканальний режим, загальний обсяг 32 ГБ) на базі програмного симулятора архітектури DRAM-Sim3. Моделювалися три типові профілі навантаження: синтетичний стрес-тест (максимальне нагрівання), робота з великими базами даних (помірне нагрівання з переважанням випадкового доступу) та офісно-мультимедійні задачі. Отримані результати вимірювань середньої латентності та пропускну здатності зведено в Таблицю 2.

Таблиця 2 – Порівняльні показники швидкодії підсистеми пам'яті при стандартному та адаптивному керуванні

Режим роботи та температурний діапазон модулів	Латентність пам'яті (Стандартний профіль JEDEC), нс	Латентність пам'яті (Запропонований Адаптивний профіль), нс	Реальний приріст продуктивності, %
Низьке навантаження (T = 35°C–42°C, офісні програми, простій)	72.4 нс	64.1 нс	+11.46 %
Змішане навантаження (T = 45°C–55°C, СУБД, компіляція ПЗ)	73.8 нс	68.5 нс	+7.18 %
Екстремальне навантаження (T = 78°C–86°C, стрес-тести, рендеринг)	78.2 нс	77.9 нс	+0.38 % (Стабільність)

Аналіз експериментальних даних з Таблиці 2 наочно демонструє, що найбільшу ефективність запропонований температурно-адаптивний комплекс виявляє в умовах низьких та помірних температур, які складають до 70 % від загального часу експлуатації середньостатистичної обчислювальної системи. Зниження латентності з 72.4 до 64.1 нс забезпечує суттєве підвищення чуйності інтерфейсу операційної системи та прискорює виконання дрібних дискових операцій [1, 2, 8]. При екстремальних навантаженнях приріст швидкодії нівелюється, проте головна мета алгоритму в даному режимі – утримання стабільності та недопущення теплового розгону системи – виконується повністю, оскільки затримки автоматично підтягуються до безпечного рівня.

Висновки. У цій роботі проведено детальний аналіз впливу температурних факторів на функціонування підсистеми оперативної пам'яті сучасних ЕОМ. Традиційні статичні методи конфігурування затримок DRAM є неоптимальними, оскільки вони змушують підсистему працювати за завищеними таймінгами, орієнтованими на критичний тепловий стан кристалів, навіть коли реальна температура модулів є безпечною. Запропонована концепція динамічного температурно-адаптивного керування дозволяє успішно реалізувати гнучке зниження таймінгів пам'яті у "холодних" зонах, що забезпечує реальний приріст швидкодії обчислювальної підсистеми в межах від 5 % до 12 % залежно від профілю навантаження.

Розроблений алгоритм гарантує високий рівень технічної безпеки: при фіксації критичного нагрівання система миттєво адаптує затримки у бік збільшення та коригує інтервали регенерації, мінімізуючи ризик пошкодження або незворотної втрати критично важливих даних користувача.

Перспективи подальших досліджень у даному напрямку полягають у розширенні математичної моделі з метою інтеграції методів машинного навчання (наприклад, використання легких штучних нейронних мереж або регресійних моделей безпосередньо у прошивці мікроконтролера керування живленням PMIC). Це дозволить реалізувати предиктивне (випереджаюче) керування не лише затримками, а й динамічною зміною напруги живлення модулів пам'яті (VDIMM) та термінальними опорами (ODT), що відкриває шлях до створення повністю автономних, енергоефективних та інтелектуальних підсистем пам'яті нового покоління.

ЛІТЕРАТУРА:

1. Стандарт JEDEC No. 79-5A. DDR5 SDRAM. JEDEC Solid State Technology Association, 2021. 414 с.
2. Стандарт JEDEC No. 79-4B. DDR4 SDRAM Specification. JEDEC Solid State Technology Association, 2017. 282 с.
3. Mutlu O. Memory Scaling: A Systems Architecture Perspective. *Proceedings of the 5th International Memory Workshop (IMW)*. Taipei, Taiwan, 2013. p. 21–25.
4. Lee K., Wong D. Thermal-aware DRAM latency management for multi-core processing units. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*. 2019. Vol. 38, No. 4. P. 612–625.
5. Hennessy J. L., Patterson D. A. Computer Architecture: A Quantitative Approach. 6th ed. Cambridge: Morgan Kaufmann, 2017. 936 с.
6. Lin X., Wang Y. Temperature-Aware Dynamic Memory Management. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2010. Vol. 18, No. 10. c. 1456–1467.
7. Mukundan J., Martinez J. F. MORPH: Flexible and Efficient Main Memory Scheduling. *IEEE Computer Architecture Letters*. 2012. Vol. 11, No. 2. c. 33–36.
8. Deshpande A. та ін. A Survey of DRAM Architectures and Performance Evaluation. *ACM Computing Surveys*. 2023. Vol. 55, No. 12. c. 1–38.

REFERENCES:

1. JEDEC Standard No. 79-5A. (2021), "DDR5 SDRAM", JEDEC Solid State Technology Association.
2. JEDEC Standard No. 79-4B (2017), "DDR4 SDRAM Specification", *JEDEC Solid State Technology Association*, 282 p.
3. Mutlu, O. (2013), "Memory Scaling: A Systems Architecture Perspective", *Proceedings of the 5th International Memory Workshop (IMW)*, Taipei, Taiwan, pp. 21–25.
4. Lee, K. and Wong, D. (2019), "Thermal-aware DRAM latency management for multi-core processing units", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 38, no. 4, pp. 612–625.
5. Hennessy J. L., Patterson D. A. Computer Architecture: A Quantitative Approach. 6th ed. Cambridge : Morgan Kaufmann, 2017. 936 с.
6. Lin X., Wang Y. Temperature-Aware Dynamic Memory Management. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2010. Vol. 18, No. 10. pp. 1456–1467.
7. Mukundan J., Martinez J. F. MORPH: Flexible and Efficient Main Memory Scheduling. *IEEE Computer Architecture Letters*. 2012. Vol. 11, No. 2. pp. 33–36.
8. Deshpande A. et al. A Survey of DRAM Architectures and Performance Evaluation. *ACM Computing Surveys*. 2023. Vol. 55, No. 12. pp. 1–38.

TEMPERATURE-ADAPTIVE MANAGEMENT OF DDR MEMORY LATENCIES IN MODERN COMPUTING SYSTEMS

Abstract. To manage modern electronic computing devices and server complexes, a critical criterion is ensuring maximum throughput of the RAM subsystem at the lowest possible latency levels. Modern dynamic random-access memory of DDR4 and DDR5 standards operates at ultra-high clock frequencies, which inevitably leads to intensive heat dissipation inside semiconductor chips under sustained workloads. An increase in the temperature of silicon crystals exponentially accelerates differential leakage currents in microscopic DRAM memory cell capacitors. This phenomenon requires a significant reduction in the intervals between forced refresh cycles (tREFI) and, as a result, leads to a significant drop in the overall performance and throughput of the computing subsystem. This review and research article examines in detail the problems of thermal throttling in modern memory profiles and proposes a comprehensive concept of dynamic temperature-adaptive management of key latencies (timings) and refresh intervals. The proposed closed-loop hardware-software system dynamically monitors the SPD thermal telemetry and adjusts major and minor timings to safely boost performance by 5–12% at low thermal ranges while ensuring data integrity under heavy processing loads.

Keywords: random-access memory, DRAM, DDR4, DDR5, memory latency, timings, temperature-adaptive control, leakage current, refresh interval, computing performance.

