

УДК 004.424:004.272.2

<http://dx.doi.org/10.30970/vam.2024.32.12383>

СУЧАСНІ ПІДХОДИ ДО РОБОТИ З НЕОДНОРІДНОЮ ПАМ'ЯТТЮ

М. Щерба, В. Білецький

Львівський національний університет імені Івана Франка,
вул. Університетська 1, Львів, 79000
e-mail: maksym.shcherba@lnu.edu.ua

Розглянуто сучасні підходи до роботи з неоднорідною пам'яттю в контексті стрімкого розвитку обчислювальних систем та їхньої енергоефективності. Проаналізовано зростання швидкодії процесорів і графічних процесорів, наведено прогноз щодо досягнення плато у їхньому розвитку. Окремо розглянуто проблему відставання пропускної здатності та затримки оперативної пам'яті від процесорної швидкодії, що створює вузькі місця в сучасних архітектурах. Висвітлено технології високопродуктивної пам'яті; пам'ять з високою пропускною спроможністю та пам'ять на основі фазового переходу. Подано перспективну технологію Compute Express Link (CXL) для розширення обсягів і пропускної здатності пам'яті, а також описано підтримку неоднорідної пам'яті з боку операційної системи Linux.

Ключові слова: неоднорідна пам'ять, пропускна здатність пам'яті, затримка пам'яті, Compute Express Link (CXL), пам'ять на основі фазового переходу (PCM), пам'ять з високою пропускною спроможністю (HBM).

1. ВСТУП

У 1965 році Гордон Мур висловив припущення, що кількість транзисторів на кристалі мікросхеми подвоюватиметься кожні два роки [12]. Це спостереження отримало назву “закон Мура”.

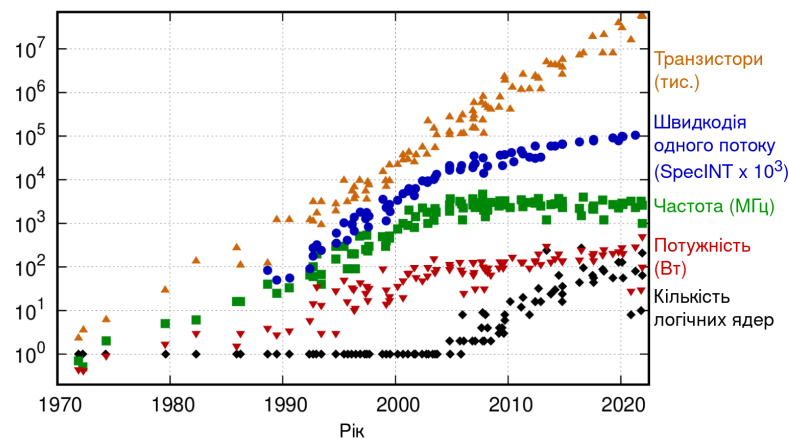


Рис. 1. Характеристики центральних процесорів за 50 років [15]
Characteristics of central processors over 50 years [15]

З графіка на рис. 1 видно, що хоч кількість транзисторів і зростає експоненційно, швидкість одного потоку та тактова частота досягли свого максимального рівня.

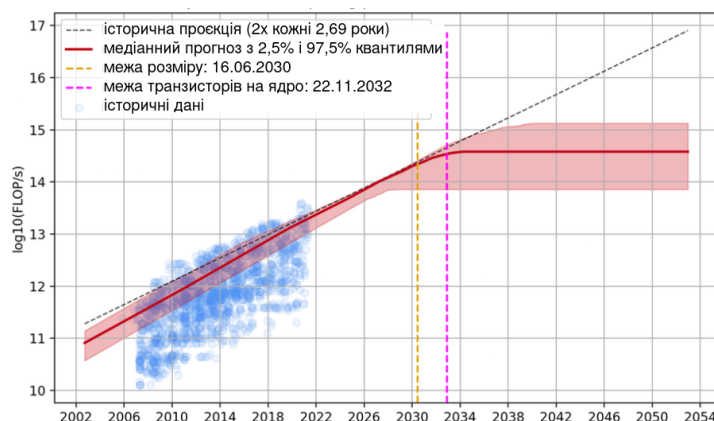


Рис. 2. Прогноз швидкодії графічних процесорів [9]
GPU performance forecast [9]

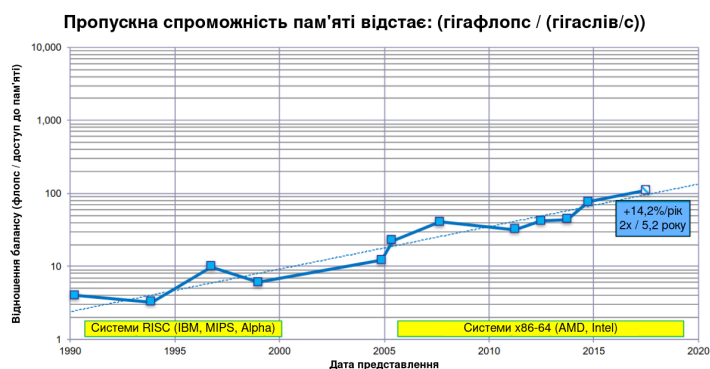


Рис. 3. Зростання швидкодії процесора порівняно з пропускнуною спроможністю пам'яті [11]
CPU speed growth versus memory bandwidth [11]

Схожою є ситуація з графічними процесорами – у [9] наведено прогноз їхньої швидкодії. Автори пропонують модель того, як швидкодія графічного процесора залежить від його характеристик, а також як характеристики змінюються з часом. Двома основними характеристиками є розмір техпроцесу та кількість ядер.

Автори зазначають, що розмір техпроцесу є наближенням до розміру транзистора. Існують різні оцінки фізичного обмеження розміру транзисторів – від 0,7 до 3 нм. У своїй моделі вони накладають це обмеження на розмір техпроцесу.

Після того, як фізична межа розміру транзистора буде досягнута, зменшення транзисторів перестане бути можливим. Однак все ще можливе збільшення кількості ядер на кристалі, що поліпшує продуктивність, забезпечуючи ефективніше паралельне обчислення. Кількість ядер продовжуватиме зростати, доки вона не досягне обмеження на кількості транзисторів, необхідних для виробництва високо-ефективних ядер. У згаданій моделі накладено обмеження на кількість транзисторів в ядрі між 600 тисячами та 2,4 мільйонами.

Результат прогнозування видно на рис. 2. Графік має дві точки залому: перша відповідає за досягнення фізичного обмеження на розмір транзистора, а друга – за той момент, коли кількість транзисторів у ядрі не можна зменшити без утрат швидкодії. Після досягнення другої межі передбачається плоска траєкторія. Модель прогнозує досягнення максимальної швидкодії від 10^{14} до 10^{15} флопс між 2027 та 2035.

На рис. 3 зображено, як зростає з часом частка від ділення швидкодії процесора на пропускну спроможність оперативної пам'яті. Ця частка зростає на 14,2% за рік. Це означає, що пропускна спроможність пам'яті відстає від швидкодії процесора.

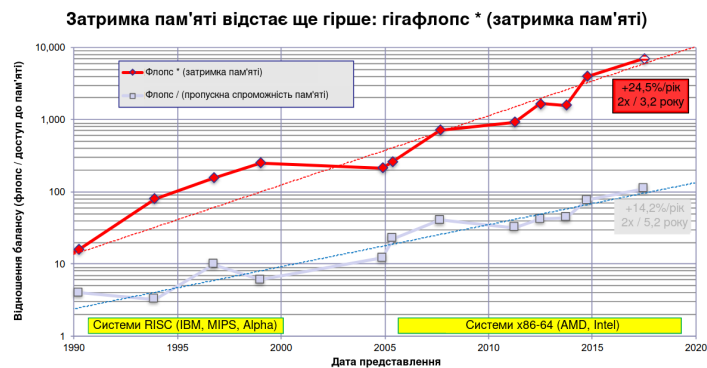


Рис. 4. Зростання швидкодії процесора порівняно з затримкою пам'яті [11]
CPU performance gain versus memory latency [11]

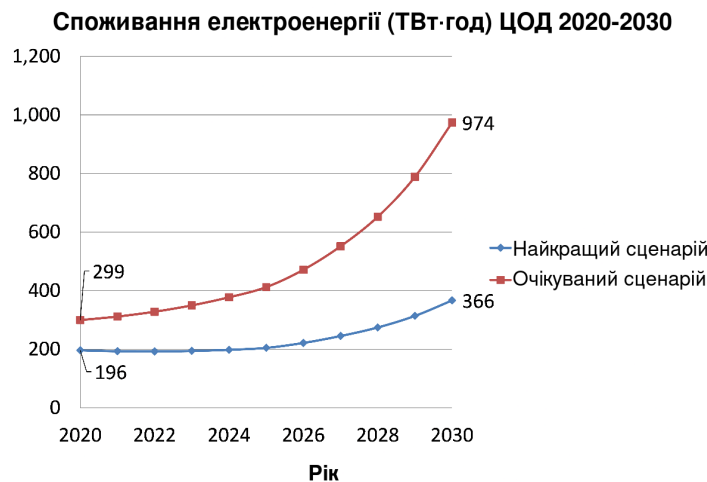


Рис. 5. Тенденція для ЦОД від 2020 до 2030 року [3]
Data Center Trends 2020 to 2030 [3]

Із затримкою пам'яті ситуація ще гірша. На рис. 4 видно, що добуток швидкодії процесора на затримку пам'яті зростає на 24,5% за рік. Тобто затримка пам'яті ще сильніше відстає від швидкодії процесора.

Окрім пропускної спроможності та затримки пам'яті, іншою важливою характеристикою пам'яті є її енергоефективність.

У [3] зроблено прогноз споживання електроенергії центрами опрацювання даних (ЦОД) протягом 2020-2030 років. Цей прогноз ґрунтується на очікуваному зростанні трафіка даних і підвищенні енергоефективності. На рис.5 зображено два сценарії: найкращий та очікуваний. У найкращому сценарії припускається дуже помірне зростання трафіка даних – у цьому разі споживання становитиме 366 ТВт·год у 2030 році.

У [5] наведено таке порівняння: одна операція із числами з рухомою комою потребує приблизно 10 пДж, тоді як один доступ до пам'яті потребує орієнтовно 10 нДж. Тобто доступ до пам'яті впотребує в ~ 1000 разів більше енергії, ніж операція з числами з рухомою комою.

Хоча кожна зі згаданих характеристик пам'яті важлива, різні застосунки мають різні потреби. Одним застосункам потрібна низька затримка, а іншим важливішою є висока пропускна спроможність. Система лише з одним типом пам'яті не задовольняє потреби всіх програм повною мірою.

Відповіддю на зазначені виклики можуть бути системи, де поєднано водночас кілька різних типів пам'яті – системи з неоднорідною пам'яттю. Далі в статті розглянуто кілька технологій пам'яті, а також засоби Linux для роботи з пам'яттю.

2. ПАМ'ЯТЬ З ВИСОКОЮ ПРОПУСКНОЮ СПРОМОЖНІСТЮ

Пам'ять з високою пропускною спроможністю (high bandwidth memory, HBM) – це тип кристала пам'яті з низьким споживанням електроенергії та великою шириною шини.

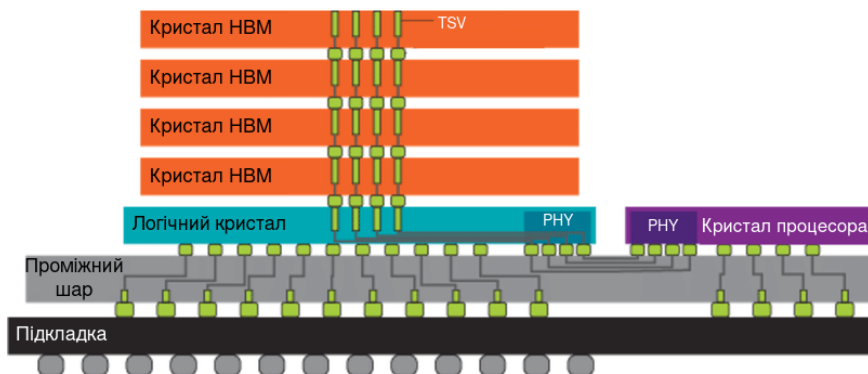


Рис. 6. Схематичне зображення HBM [10]
Schematic representation of HBM [10]

Кристали пам'яті складають вертикально один на одного. З'єднання між кристалами називаються TSV (through-silicon vias) (рис. 6).

Міжз'єднання TSV формуються шляхом витравлювання перехідних отворів у шарах кристалів, що з'єднуються (рис. 7). Отвори заповнюються струмопровідним металом. [2]

Першою широкодоступною системою для HPC з неоднорідною пам'яттю була

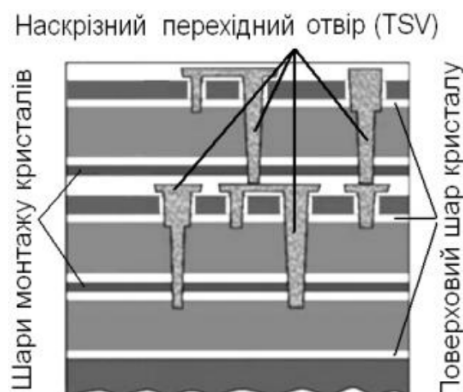


Рис. 7. Схематичне зображення TSV [2]
Schematic representation of TSV [2]

Intel *Knights Landing* Xeon Phi (KNL), випущена у 2016 році. У цей багатоядерний процесор була вбудована багатоканальна динамічна оперативна пам'ять (dynamic RAM, DRAM) з пропускною здатністю 400 ГБ/с, тоді як більша зовнішня DRAM досягала лише близько 100 ГБ/с. Багатоканальну DRAM можна було використовувати як окрему область пам'яті, що потребувало робити вибір між малим обсягом і високою пропускною спроможністю, з одного боку, та великим обсягом і низькою пропускною спроможністю, з іншого. Також багатоканальна DRAM могла слугувати кешем перед звичайною DRAM. Такий режим *cache* жертвує швидкістю заради простоти, оскільки не потрібно змінювати застосунки, щоб вони явно виділяли потрібний їм тип пам'яті.

Зараз випуск серії KNL припинено і донедавна в HPC не було подібних підсистем пам'яті. Однак деякі виробники процесорів оголосили повернення інтегрованої пам'яті з високою пропускною здатністю. Наприклад, Intel Xeon *Sapphire Rapids* має вбудовані 64 ГБ HBM. У поєднанні зі звичайною зовнішньою DRAM ця підсистема пам'яті дуже схожа до KNL. [8]

3. ПАМ'ЯТЬ НА ОСНОВІ ФАЗОВОГО ПЕРЕХОДУ

Пам'ять на основі фазового переходу (phase-change memory, PCM) – це тип енергонезалежної пам'яті, що використовує властивість халькогенідного скла переходити між двома станами, аморфним і кристалічним, під впливом тепла, що генерується електричними імпульсами. Халькогенідне скло можна надійно, швидко й багаторазово перемикає з однієї фази на іншу. Аморфна фаза має високий електричний опір, а кристалічна – низький. Різниця в опорі двох станів становить приблизно п'ять порядків, і може використовуватися для визначення логічних станів бінарних даних.

Хоча принцип використання зміннофазних матеріалів для комірки пам'яті був продемонстрований у 1960-х роках [13], технологія була надто повільною, щоб бути практичною. Однак відкриття матеріалів, що швидко кристалізуються, таких як $\text{Ge}_2\text{Sb}_2\text{Te}_5$ (GST) і легованого сріблом та індієм Sb_2Te (AIST), відновило промис-

ловий інтерес до PCM. І GST, і AIST можуть кристалізуватися менш ніж за 100 нс порівняно з понад 10 мкс для попередніх матеріалів. Було виготовлено та випробувано пристрої PCM із надзвичайно малими розмірами – 3 нм × 20 нм [14].

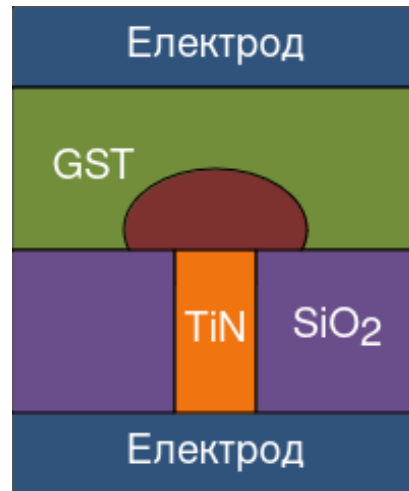


Рис. 8. Будова комірки PCM [1]
Structure of a PCM cell [1]

На рис. 8 зображено базову структуру пристрою пам'яті на основі фазового переходу. GST розташований між верхнім і нижнім електродами з нагрівальним елементом (з TiN), який виходить від нижнього електрода та налагоджує контакт з матеріалом.

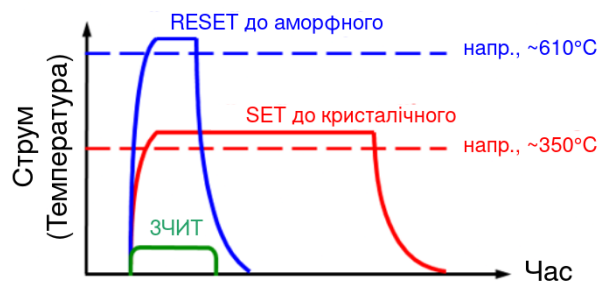


Рис. 9. Принцип роботи комірки PCM [1]
The principle of operation of the PCM cell [1]

Опишемо детальніше процеси, які відбуваються у комірці PCM під час роботи. Спочатку сплав перебуває в аморфному стані, тобто має високий опір. Якщо прикладена напруга перевищує певне порогове значення, то починає протікати струм, який розігріває сплав.

Якщо протягом заданого проміжку часу (наприклад, 100 нс) підтримувати температуру нагрівання, яка в цьому випадку становить приблизно 350 °C, то сформується кристалічна ґратка, яка забезпечує низьке значення електричного

опору. У цьому випадку кристалічна структура не руйнується після того, як напруга знімається і сплав охолоджується. Така кристалізація змінює стан біта з логічного “0” на логічну “1” і називається операцією SET.

Тепер для того, щоб прочитати значення, що зберігається в комірці біта, а по суті, виміряти опір, не потрібні високі напруги й струми. Досить прикласти невелику напругу, скажімо, 0,5 В, щоб отримати струм порядку 0,5 мА – логічну “1”.

Висока напруга і, відповідно, великий струм, знадобляться, коли треба буде перевести сплав назад в аморфний стан. У наведеному прикладі потрібна температура близько 600 °С, яка призводить до розплавлення матеріалу. Після зняття напруги сплав швидко охолоджується, у цьому випадку виявляється, що при проході через область кристалізації, часу для формування будь-якої кристалічної структури не вистачає. Тут значення біта змінюється з “1” на “0”. Така операція називається RESET.

Тепер, якщо прикласти ту саму напругу в 0,5 В, високий опір аморфного сплаву призводить до того, що значення струму практично дорівнює нулю. Отож отримуємо інше значення біта, що зберігається в комірці – це логічний “0” [1].

Цю технологію пам’яті можна використовувати для постійного зберігання, а також як основну пам’ять великого обсягу. Тоді DRAM є швидкою пам’яттю малого обсягу, на відміну від KNL, де DRAM розглядається як повільна пам’ять великого обсягу [14].

4. COMPUTE EXPRESS LINK (CXL)

Compute Express Link (CXL) – це відкритий стандарт, розроблений спільними зусиллями основних виробників апаратного забезпечення, для високої пропускної спроможності та низької затримки при робочих навантаженнях у ЦОД. Цей стандарт визначає протоколи з’єднання між головним процесором і пристроями такими, як прискорювачі, буфери оперативної пам’яті та пристрої вводу-виводу. [6]

У стандарті CXL визначено три окремі протоколи: CXL.io, CXL.cache і CXL.memory.

- Протокол CXL.io побудований на основі інтерфейсу PCI Express (PCIe) і використовується для таких функцій, як виявлення пристроїв, конфігурація, ініціалізація, віртуалізація вводу-виводу та прямий доступ до пам’яті з використанням некогерентного завантаження та запам’ятовування.
- CXL.cache дає змогу пристрою кешувати дані з головної пам’яті, застосовуючи простий протокол запиту й відповіді. Головний процесор керує когерентністю розміщених у кеші пристрою даних.
- CXL.memory дає змогу головному процесорові доступатися до пам’яті пристрою CXL. Транзакції CXL.memory – це прості транзакції завантаження та запам’ятовування, що йдуть від головного процесора, який дбає про когерентність [6].

Оскільки CXL є надбудовою над стандартом PCIe, то він пропонує набагато вищу швидкість передавання (наприклад, PCIe 4.0: 16 Гбіт/с проти DDR4-3200: 3,2 Гбіт/с) і потребує менше енергії на передавання (наприклад, PCIe 4.0: 6 пДж/біт проти DDR4: 22 пДж/біт), але коштом набагато довшої затримки (наприклад, PCIe 4.0: 40 нс проти DDR4: <1 нс). Пристрій пам’яті CXL на основі PCIe 5.0 може розширити обсяг пам’яті та пропускну здатність систем. Крім того, з

контролером CXL між процесором і пристроями пам'яті CXL розмежовує технології пам'яті від підтримуваної процесором конкретної технології інтерфейсу пам'яті. Це дає виробникам пам'яті безпрецедентну гнучкість у розробці й оптимізації їхніх пристроїв пам'яті [7].

Таблиця 1

Пропускні спроможності, затримки та обсяги різних видів пам'яті.
Усі ці значення дуже приблизні, оскільки вони залежать від конкретної моделі процесора, кількості каналів пам'яті, частот і т.д. [8]
Bandwidths, latencies, and capacities of different types of memory.
All of these values are very approximate, as they depend on the specific processor model, number of memory channels, frequencies, etc. [8]

Технологія	Пропускна спроможність, (ГБ/с)	Затримка, (нс)	Обсяг, (ГБ)
DRAM	200	100	100
Багатоканальна DRAM	400	120	16
HBM2	800	120	64
Енергонезалежна пам'ять	50	300	1000
CXL DRAM	50	250	1000

Вищезгадані види пам'яті мають значні розбіжності у швидкодії пам'яті: багатоканальна DRAM і HBM мають набагато вищу пропускну спроможність, ніж DRAM, але їхня затримка часто є вищою. Енергонезалежна пам'ять має суттєво вищу затримку та нижчу пропускну спроможність, але її обсяг більший. Також обсяг зазвичай зменшується зі зростанням пропускну спроможності, але не корелює із затримкою. У табл. 1 підсумовано приблизні характеристики цих технологій [8].

5. ЗАСОБИ LINUX ДЛЯ РОБОТИ З ПАМ'ЯТТЮ

Процеси Linux можуть використовувати функцію `mmap`, щоб створювати нові ділянки віртуальної пам'яті й відображати об'єкти в ці ділянки.

```
void *mmap(void *start, size_t length,
           int prot, int flags, int fd, off_t offset);
```

Функція `mmap` робить запит до ядра створити нову ділянку віртуальної пам'яті, яка бажано починається за адресою `start`, та відобразити неперервний шматок об'єкта, заданого дескриптором файлу `fd`, в нову ділянку. Неперервний шматок об'єкта має розмір `length` байтів та розпочинається за `offset` байтів від початку файлу. Адреса `start` є лише підказкою і зазвичай зазначається як `NULL`.

Аргумент `prot` містить біти, що описують дозволи доступу ділянки віртуальної пам'яті.

`PROT_EXEC`. Ділянка складається з інструкцій, які може виконувати центральний процесор.

`PROT_READ`. З ділянки можна зчитувати.

`PROT_WRITE`. У ділянку можна записувати.

`PROT_NONE`. До ділянки не можна доступатися.

Аргумент `flags` складається з бітів, що описують тип відображеного об'єкта.

Функція `munmap` видаляє області віртуальної пам'яті:

```
int munmap(void *start, size_t length);
```

Функція `munmap` видаляє ділянку, що розпочинається за віртуальною адресою `start` і складається з наступних `length` байтів. Наступні звертання до видаленої області спричинять помилку.

Попри те, що безсумнівно можна використовувати низькорівневі функції `mmap` і `munmap`, щоб створювати та видаляти ділянки віртуальної пам'яті, зручніше використовувати динамічний розподільник пам'яті.

Стандартна бібліотека C надає розподільник, відомий як пакет `malloc`. Програми виділяють блоки пам'яті викликом функції `malloc`.

```
void *malloc(size_t size);
```

Функція `malloc` повертає покажчик на блок пам'яті розміром принаймні `size` байтів. Якщо `malloc` нашоїхнеться на проблему (наприклад, програма спробувала виділити блок пам'яті, більший, ніж доступна віртуальна пам'ять), то вона поверне `NULL`.

Програми звільняють виділені блоки пам'яті викликом функції `free`.

```
void free(void *ptr);
```

Аргумент `ptr` повинен показувати на початок виділеного блока [4].

У випадку неоднорідної пам'яті користувач може зазначити, який саме тип пам'яті йому потрібен. Якщо різні типи пам'яті розглядаються як різні вузли NUMA (non-uniform memory access), то можна викликати функцію `mbind`.

```
int mbind(void *start, unsigned long length, int mode,  
          unsigned long *nodemask, unsigned long maxnode, unsigned flags);
```

`mbind` встановлює політику пам'яті NUMA для проміжку пам'яті, що розпочинається за адресою `start` і має розмір `length` байтів.

`nodemask` показує на бітову маску вузлів NUMA.

6. ВИСНОВКИ

Було розглянуто сучасні підходи до роботи з неоднорідною пам'яттю в контексті розвитку обчислювальних систем та їхньої енергоефективності. Огляд технологій пам'яті продемонстрував, що хоча процесори та графічні процесори продовжують швидко розвиватися, швидкість розвитку технологій оперативної пам'яті значно відстає, що створює вузькі місця в обчислювальних архітектурах.

Особливо актуальною стала проблема розриву між швидкістю процесорів та пропускнуною спроможністю і затримкою пам'яті, що впливає на загальну продуктивність систем. Аналіз різних видів пам'яті, таких як пам'ять з високою пропускнуною здатністю (HBM), пам'ять на основі фазового переходу (PCM) та технологія Compute Express Link (CXL), продемонстрував потенціал цих рішень у подоланні зазначених обмежень.

Зокрема, НВМ, з її високою пропускною спроможністю та енергоефективністю, є перспективною для використання в системах високопродуктивних обчислень (НРС). РСМ пропонує можливість створення пам'яті, яка здатна зберігати дані навіть без живлення, що особливо цінно для постійного зберігання великих обсягів даних. Технологія CXL відкриває нові можливості для розширення пам'яті та підвищення її пропускної здатності, що може суттєво знизити затримки в системах із високими вимогами до обробки даних.

Розглянуто також підтримку операційною системою Linux роботи з неоднорідною пам'яттю, що є важливим для забезпечення ефективного використання нових видів пам'яті на рівні програмного забезпечення. Така підтримка допомагає краще використовувати доступні ресурси та мінімізувати негативний вплив затримок пам'яті.

Отже, майбутнє комп'ютерних систем потребує впровадження нових підходів до роботи з пам'яттю, які зможуть забезпечити збалансований розвиток обчислювальної потужності та пропускної здатності пам'яті, знижуючи енергоспоживання та підвищуючи загальну ефективність систем. Це є критичним для підтримки чимраз більших потреб у розробці сучасних обчислювальних систем та обробці великих обсягів даних.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Залевський Д.В. Властивості матеріалів для резистивної пам'яті з довільним доступом: дис. канд. фіз.-мат. наук: 104 / Залевський Денис Віталійович. – Кривий Ріг, 2023. – 169 с.
2. Філяшкін М.К. Мікроелектромеханічні системи / Микола Кирилович Філяшкін. – Київ: НАУ, 2019. – 276 с.
3. Andrae A. New perspectives on internet electricity use in 2030 / Anders S.G. Andrae // Engineering and Applied Science Letters. – 2020. – № 2. – С. 19–31.
4. Bryant R. Computer Systems: A Programmer's Perspective / R. Bryant, D. O'Hallaron. – 2016.
5. Dally B. Challenges for Future Computing Systems [Електронний ресурс] / Bill Dally // HiPEAC. – 2015. – Режим доступу до ресурсу: <https://www.cs.colostate.edu/cs575dl/Sp2015/Lectures/Dally2015.pdf>.
6. Das Sharma D. Compute Express Link [Електронний ресурс] / Debendra Das Sharma // CXL Consortium White Paper. – 2019. – Режим доступу до ресурсу: https://docs.wixstatic.com/ugd/0c1418_d9878707bbb7427786b70c3c91d5fbd1.pdf.
7. Sun Y. Demystifying CXL Memory with Genuine CXL-Ready Systems and Devices / Y. Sun, Y. Yuan, Z. Yu та ін. // Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture. – 2023. – С. 105–121.
8. Foyer C. A survey of software techniques to emulate heterogeneous memory systems in high-performance computing / C. Foyer, B. Goglin, A. Rubio Proano // Parallel Comput. – 2023.
9. Hobbhahn M. Predicting GPU Performance [Електронний ресурс] / M. Hobbhahn, T. Besiroglu. – 2022. – Режим доступу до ресурсу: <https://epochai.org/blog/predicting-gpu-performance>.
10. Macri J. AMD's next generation GPU and high bandwidth memory architecture: FURY / Joe Macri // IEEE hot chips 27 symposium (HCS). – 2015. – С. 1–26.
11. McCalpin J. Memory Bandwidth and System Balance in HPC Systems [Електронний ресурс] / John David McCalpin // Invited Talk at the Supercomputing. – 2016. – Режим доступу до ресурсу: <https://repositories.lib.utexas.edu/items/94632694-ae6a-4521-8a73-0d833e869db3>.

12. Moore G. Cramming More Components onto Integrated Circuits / Gordon E. Moore // Electronics. – 1965. – № 8. – С. 114–117.
13. Ovshinsky S. Reversible Electrical Switching Phenomena in Disordered Structures / Stanford R. Ovshinsky // Phys. Rev. Lett. – 1968. – № 20. – С. 1450–1453.
14. Qureshi M. Scalable high performance main memory system using phase-change memory technology / M. Qureshi, V. Srinivasan, J. Rivers // SIGARCH Comput. Archit. News. – 2009. – № 3. – С. 24–33.
15. Rupp K. Microprocessor Trend Data [Електронний ресурс] / Karl Rupp. – Режим доступу до ресурсу: <https://github.com/karlrupp/microprocessor-trend-data>.

Стаття: надійшла до редколегії 10.10.2024

доопрацьована 25.10.2024

прийнята до друку 14.11.2024

MODERN APPROACHES TO WORKING WITH HETEROGENEOUS MEMORY

M. Shcherba, V. Biletskyy

*Ivan Franko National University of Lviv,
1, Universytetska str., 79000, Lviv, Ukraine
e-mail: maksym.shcherba@lnu.edu.ua*

The article explores modern approaches to working with heterogeneous memory in the context of the rapid development of computing systems and their energy efficiency. The growth of CPU and GPU performance is analyzed, and a forecast is provided regarding the plateau expected in their development. The issue of memory bandwidth and latency lagging behind CPU performance is highlighted, creating bottlenecks in modern architectures. High-performance memory technologies such as high bandwidth memory (HBM) and phase-change memory (PCM) are examined. The article also introduces the promising Compute Express Link (CXL) technology for expanding memory capacity and bandwidth. Additionally, the support for heterogeneous memory in the Linux operating system is described.

Key words: heterogeneous memory, memory bandwidth, memory latency, Compute Express Link (CXL), phase-change memory (PCM), high bandwidth memory (HBM).