На пути к «чистой» архитектуре СУБД на основе энергонезависимой основной памяти

С.Д. Кузнецов

Институт системного программирования им. В.П. Иванникова РАН

Введение (1)

- Энергонезависимая основная память
  - Non-Volatile Main Memory, NVM,
  - Persistent Main Memory,
  - Storage-Class Memory
- становится все более реальной
- К массовому выпуску соответствующих чипов приступили Intel и Samsung
- Расширяется набор технологий, используемых даже внутри одной компании для производства чипов NVM
- Однако число исследователей в области NVM-ориентированных СУБД за последние два-три года не увеличилось

Введение (2)

- Выделяются два молодых исследователя:
- Джой Арулрадж (Joy Arulraj) : PhD в 2018 г. в университете Карнеги-Меллон
  - Под руководством Энди Павло (Andy Pavlo, Pavlopolis?)
- Премия им. Джима Грея за лучшую диссертационную работу на SIGMOD 2019
- С 2018 г. в Технологическом институте Джорджии
- Исмаил Укид (Ismail Oukid): PhD [8] в 2017 г. в Дрезденском техническом университете
  - Под руководством Вольфганга Лехнера (Wolfgang Lehner)
- С 2018 г. в SAP

Введение (4)

- Сравнительно равнодушное отношение старшего поколения сообщества баз данных к теме NVM-ориентированных СУБД
- Данные о реальных характеристиках промышленных образцов MVM недостаточны и противоречивы
- Вендоры NVM на первых порах предпочитают предлагать продукты в форм-факторе блочных устройств внешней памяти
- Трудно получить гранты (большие риски)
- Старшие члены сообщества баз данных на самом деле в своих исследованиях затрагивают вопросы, прямо относящиеся к архитектуре и алгоритмам NVM-ориентированных СУБД
  - In-memory СУБД

Введение (5)

- Никто не решается замахнуться на полноценную разработку СУБД с одноуровневой системой хранения
- С одной стороны, расчет на использование иерархически организованной системы хранения с использованием NVM ближе к сегодняшней реальности
- С другой стороны, именно полный переход к одноуровневой энергонезависимой системе хранения может позволить упростить организацию СУБД и значительно увеличить их эффективность

Аппаратные средства NVM (1)

- Как и раньше, три основных технологии
- Фазовые переходы (Phase-Change Memory, PCM),
- Изменение сопротивления диэлектриков (Resistive Random-Access Memory, ReRAM),
- Перенос спинового момента (Spin-Torque-Transfer Memory, STT-RAM)
  - STT-RAM относится к более общему классу магнитно-резистивных решений (magnetoresistive random-access memory, MRAM)
  - Все разновидности MRAM, как и ReRAM, основаны на изменении сопротивления
  - В случае MRAM сопротивление изменяется при изменении ориентации намагниченности

Аппаратные средства NVM (2)

- Как и раньше, три основных технологии
- Ближе всех к DRAM по скорости находится STT-RAM

	Год публикации	PCM	STT-RAM	ReRAM	DRAM
Время чтения (нс)	2019	20-70	10-30	10	10-50
	2018	250	20	100?	100
	2017	50	20	100	60
	2016	50	10	10	50
Время записи (нс)	2019	50-500	13-95	1-100	10-50
	2018	250	20	100?	100
	2017	150	20	100	60
	2016	500	50	50	50
Стойкость	2019 *	$1\text{-}10^8$	10^{15}	$10^{10}\text{-}10^{12}$	$> 10^{17}$
	2018 **	100	↑	40?	↑
	2017 *	10^{10}	10^{15}	10^8	$> 10^{16}$
	2016 *	$10^8\text{-}10^9$	$> 10^{15}$	10^{11}	$> 10^{15}$
*** Число циклов записи**					
**** DWPD, Drive Writes Per Day, допустимый объем суточной записи относительно емкости самого устройства**					

Андрей Николаенко, IBS, Tarantool Conference, 2018

Аппаратные средства NVM (3)

- PCM: Micron Technology
- ReRAM: Panasonic и Crossbar
- STT-RAM: Everspin Technologies, Crocus Technology, Крокус Наноэлектроника
- Ни одна из компаний не является вендором мирового уровня, способным (или хотя бы желающим) производить доступные для широкого использования модули энергонезависимой памяти
- Реальным прорывом является начало массового производства продуктов категории NVM компаниями Intel и Samsung

Аппаратные средства NVM (4)

- 2015 г.: Intel и Micron объявили о создании технологии 3D XPoint – NVM на основе PCM
- 2017 г.: Intel объявила о промышленном производстве модулей 3D XPoint под брендом Optane
  - по скоростным характеристикам NVM Optane занимает нишу между DRAM и флэш-памятью
  - кэш внутри традиционных дисковых устройств и
  - твердотельные диски целиком на основе Optane

Аппаратные средства NVM (5)

- Август 2018 г.: Intel объявила о начале производства DIMM на основе Optane
  - невозможно использовать с обычными процессорами Intel
  - новая линейка процессоров Cascade Lake AP Xeon CPU
  - использовать Optane NVM как замену RAM проблематично
- В начале 2018 г. Samsung объявила о начале производства SSD на основе собственной технологии Z-NAND
  - повысили скорость чтения в 10 раз по сравнению с обычными SSD
  - конкурентоспособные с Optane SSD результаты
- Стоит еще раз задуматься об архитектуре СУБД на основе SSD

Аппаратные средства NVM (6)

- В феврале 2019 г. Intel заявила о готовности начать промышленный выпуск модулей памяти STT-MRAM
  - Solid-State Circuits conference
- Там же было заявлено о готовности Intel к разработке моделей памяти и на основе технологии ReRAM
  - Интернет вещей и автомобилестроение
- Через две недели Samsung объявила о начале поставок модулей энергонезависимой памяти eMRAM на основе STT
  - пока модули eMRAM будут использоваться в Интернете вещей
  - но если технология успешно себя зарекомендует, то Samsung сможет обеспечить и выпуск моделей памяти в формате DIMM
- Высокопроизводительная энергонезависимая основная память стала реальностью

Родственные работы (1)

- Архитектурные и алгоритмические черты in-NVM СУБД
- Работ мало, хотя тематике NVM посвящено много статей
- Стратис Виглас, 2015 (Stratis D. Viglas, Эдинбургский университет): два пути интеграции NVM в иерархию сред хранения данных: путем использования в качестве устройства внешней памяти и на основе подключения таким же способом, что и основная память
- Навид Уль Мустафа и др., 2016 (Naveed Ul Mustafa, университет Билькент, Анкара, Барселонский суперкомпьютерный центр): замена в PostgreSQL дисковой подстистемы хранения на блочную подсистему на основе NVM
  - через файлы, отображаемые в память

Родственные работы (2)

- Михня Андрей и др., 2017 (Mihnea Andrei, SAP SE, German division of Intel): первая попытка внедрить использование NVM в СУБД HANA; в NVM-DIMM размещаются лишь редко изменяемые крупные фрагменты описателей таблиц
- Александр ван Ренен и др., 2018 (Alexander van Renen, Мюнхенский технический университет): архитектура системы хранения данных, в иерархии которой NVM (3D Xpoint) занимает место между DRAM и SSD; «решение, конкурентоспособное по отношению к in-memory и in-NVM СУБД»
  - «Не позволяйте себя дурачить книгам о сложности или всяким сложным и малопонятным алгоритмам, которые вы найдете в этой книге или где-то еще. Хотя нет учебников по простоте, простые системы работают, а сложные нет» (Джим Грей, Transaction Processing: Concepts and Techniques)

Родственные работы (3)

- Джой Арулрадж: проект Peloton, лидер - Энди Павло
- Peloton – «самоуправляемая» (self-driving) in-memory СУБД
- Поведение системы контролируются интегрированным планировщиком, который оптимизирует систему для выполнения текущей рабочей нагрузки и прогнозирует будущую рабочую нагрузку, чтобы система к ней подготовиться
- Кроме этого, в Peloton реализована поддержка энергонезависимой основной памяти

Родственные работы (4)

- В архитектуре подсистемы хранения данных используются свойства долговечности и байтовой адресации NVM; обеспечивается экономное использование энергонезависимой памяти и увеличивается срок ее эксплуатации за счет уменьшения числа записей
- В ней применяется новый протокол журнализации и восстановления «журнализация с отложенной записью в журнал» (Write-Behind Logging, WBL), позволяющий обеспечить высокий уровень доступности

Родственные работы (5)

- Исмаил Укид, Дрезденская группа систем баз данных
- Прототип NVM-ориентированной СУБД SOFORT
- СУБД категории HTAP (Hybrid Transactional and Analytical Processing)
  - кстати, Peloton тоже HTAB
- Тесные связи Дрезденской группы вообще и ее руководителя Лехнера с проектом гибридной поколоночной in-memory СУБД HANA компании SAP
- SOFORT – СУБД с одноуровневой системой хранения, использующей как NVM, так и традиционную энергозависимую основную память

Родственные работы (6)

- Результаты Укида:
- модель программирования с использованием NVM;
- методы управления энергонезависимой памятью и ее распределения для нужд СУБД;
- методы управления транзакциями и восстановления баз данных после сбоев;
  - адаптирован, оптимизирован и реализован многоверсионный оптимистический протокол
- разработан фреймворк для тестирования NVM-ориентированного программного обеспечения

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (1)

- Инфраструктурные ограничивающие факторы
- Недостаточная развитость аппаратных технологий NVM и/или недостаточная информированность сообщества баз данных со стороны вендоров
- Влияние решений, наиболее активно используемых в близкой, но принципиально отличающейся области in-memory СУБД

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (2)

- **Инфраструктурные ограничивающие факторы**
- Инфраструктура должна включать подготовленных и заинтересованных специалистов, вместе с которыми можно было бы вырабатывать решения
- Инфраструктура должна обеспечивать возможность быстрого построения прототипов системы, поддерживать необходимые процессы тестирования и отладки
- Необходимо наличие как готовых для использования аппаратно-программных инструментальных средств, так и программистов, способных быстро и достаточно качественно участвовать в разработке
- Инфраструктура должна помогать находить источники финансирования новых проектов

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (3)

- Лучше всего инфраструктуру обеспечивают исследовательские лаборатории компаний или университетов
- Примерами являются группа баз данных в университете Карнеги-Меллон и Дрезденская группа систем баз данных
- Но трудно решиться на открытие совсем нового проекта по теме, которая не гарантирует абсолютного успеха
- Поэтому естественно желание «пристегнуть» новый проект к существующему, зарекомендовавшему себя и устойчиво финансируемому проекту: возникают ограничения
- «Чистую» архитектуру одноуровневой СУБД можно разработать только с нуля, не подвергаясь влиянию особенностей инфраструктуры

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (4)

- **Ограничивающие предположения о характеристиках аппаратных средств NVM**
- Исходя из недостаточно достоверных сведений, исследователи принимают разные пессимистические предположения о характеристиках ожидаемой энергонезависимой памяти.
- Некоторые полагают, что NVM является и навсегда останется существенно более медленной, чем DRAM, и поэтому следует использовать NVM в качестве еще одного слоя в иерархии систем хранения данных между DRAM и DDR
- Другие считают, что неустранимой особенностью всех видов байт-адресуемой NVM является существенная разница в скорости выполнения операций чтения и записи: запись медленнее чтения
- Наконец, еще одним распространенным предположением является меньшая стойкость (endurance) NVM по сравнению как с DRAM, так и с DDR

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (5)

- Могут оказаться востребованными различные архитектуры СУБД с использованием NVM
- При наличии не слишком быстрой, но достаточно стойкой NVM может оказаться достаточно эффективной многоуровневая организация системы хранения DRAM-NVM
- При аналогичных показателях NVM вполне перспективен и путь к использованию иерархии DRAM-NVM-DDR (или SSD)
- Еще один путь к использованию NVM в форм-факторе DIMM состоит в том, чтобы вообще не использовать байт-адресуемую энергонезависимую память в архитектуре будущей СУБД
  - быстрые SSD

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (6)

- Нас интересует чистая одноуровневая архитектура NVM-ориентированной СУБД, в которой применяются только байт-адресуемые DRAM и NVM
- Чтобы обеспечить «чистоту» (nativeness) этой архитектуры, мы должны опираться на оптимистические предположения относительно характеристик NVM:
  - скорость произвольного доступа к NVM одинакова для операций чтения и записи и не уступает скорости доступа к DRAM
  - стойкость NVM не уступает стойкости DDR (SSD)
- Возможно, предположения слишком оптимистичны, но они соответствуют наблюдаемым тенденциями развития технологии NVM

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (7)

- **Влияние методов, используемых в близких областях**
- NVM обладает важным отличием от устройств внешней памяти:
  - это байт-адресуемая память прямого доступа
- NVM принципиально отличается и от традиционной основной памяти:
  - она энергонезависима и потому обеспечивает долговременное хранение данных
- Можно найти много примеров необоснованных заимствований решений; остановимся на двух, относящимся к важным вопросам индексации и управления транзакциями

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (8)

- ***Индексация***
- Для организации индексов в дисковых СУБД наибольшей популярностью пользуются B+-деревья
- Сильная ветвистость и полная сбалансированность B+-дерева играют на пользу индексации, если индекс хранится во внешней памяти
- При хранении B+-деревьев в основной памяти наличие отдельной поисковой структуры приводит только к дополнительным накладным расходам памяти и процессорных ресурсов
- Тем не менее, изобретено большое число разновидностей B+-деревьев в основной памяти, и они активно заимствуются в новые архитектуры NVM-ориентированных СУБД в блочной форме

Факторы, ограничивающие «чистоту» архитектуры in-NVM СУБД (9)

- ***Управление транзакциями***
- Во всех современных in-memory СУБД используется какой-либо вариант протокола MVCC
  - MultiVersion Timestamp Ordering
  - MVTO Two-Version 2PL, 2V2PL
  - MultiVersion Optimistic Concurrency Control, MVOCC
- Идеальных решений нет
- Специфические черты конкретного механизма управления транзакциями накладывают отпечаток на многие другие компоненты СУБД
- Выбору уделяется недостаточное внимание

Набросок архитектуры «чистой» NVM-ориентированной СУБД (1)

- **Предположения**
- NVM обладает всеми «оптимистическими» характеристиками
- Целевой компьютер содержит требуемое число ядер или потоков управления, поддерживаемых аппаратурой, а также достаточные объемы энергозависимой и энергонезависимой основной памяти
- Внешняя память совсем не используется.
- Вся серверная часть системы выполняется на одном компьютере, в энергонезависимой основной памяти которого сохраняется баз данных
- Архитектура рассчитана на поддержку только транзакционных рабочих нагрузок.
- При использовании аппаратных средств не допускается никакая виртуализация
  - СУБД сама следит за распределением энергонезависимой основной памяти
  - Каждой транзакции жестко соответствует отдельное ядро процессора или поток управления, поддерживаемый аппаратурой

Набросок архитектуры «чистой» NVM-ориентированной СУБД (2)

- **Распределение энергонезависимой памяти**
- Предлагается основывать распределение памяти на классическом методе двоичных близнецов
  - размер динамически запрашиваемой энергонезависимой в in-NVM СУБД всегда $\geq 2^6$, а в основном будут запрашиваться участки памяти именно этого размера
  - чтобы избежать потребности в синхронизации параллельно выполняемых потоков управления, свободная энергонезависимая память заранее делится поровну между всеми рабочими потоками управления, и в каждом потоке работает собственный компонент распределения памяти в своей части общей кучи
- Потребуются и неявные запросы памяти («по требованию») при расширении сегментов, требуемых для поддержки индексов
  - придется влезать в ядро ОС (Linux?)

Набросок архитектуры «чистой» NVM-ориентированной СУБД (3)

- **Хранение таблиц и управление транзакциями**
- Предлагается использовать в «чистой» архитектуре NVM-ориентированной СУБД разновидность протокола 2V2PL
- 2V2PL предполагает хранение двух копий каждой строки каждой таблицы базы данных:
  - текущей версии, созданной последней зафиксированной транзакцией, и
  - измененной версии, созданной еще не зафиксированной транзакцией
- Естественно, это влияет на организацию хранения таблиц в NVM

Набросок архитектуры «чистой» NVM-ориентированной СУБД (4)

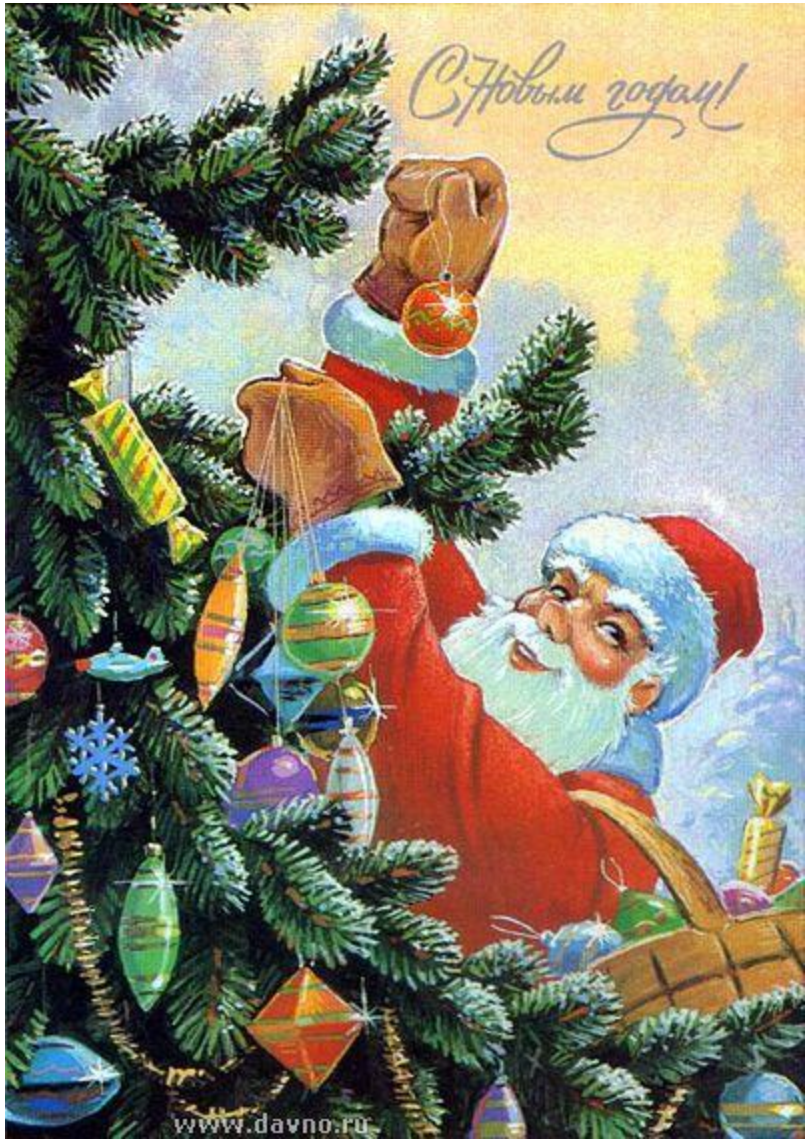
- **Организация индексов**
- Считаем реалистичным бенчмарк TPC-C
- В транзакционной NVM-ориентированной СУБД требуются индексы, обеспечивающие быстрый прямой доступ к строкам таблиц по значению указанного столбца
  - наиболее просто и эффективно можно реализовать индекс на основе варианта метода линейного хеширования
- Для поиска по диапазону значений ключа придется обеспечивать другой вид индексов, скорее всего, на основе B-деревьев
  - в первом варианте все узлы дерева представляют списки бакетов по 64 байта, а критерием расщепления или слияния соседних узлов является длина списка
  - во втором варианте каждое B-дерево располагается в отдельном расширяемом сегменте виртуальной памяти, а узел дерева – линейный список элементов, располагающийся в пределах одной страницы сегмента

Набросок архитектуры «чистой» NVM-ориентированной СУБД (5)

- **Оптимизация запросов: три замечания**
- Аспекты оптимизации SQL-запросов в NVM-ориентированных СУБД к настоящему времени исследованы явно недостаточно
- Традиционным способом организации транзакционных приложений баз данных является перенос всей логики приложения на сторону сервера баз данных обычно в виде хранимых процедур
  - компиляция и оптимизация декларативных оператором SQL должна производиться тогда же, когда выполняется компиляция кода хранимой процедуры
  - должна производиться полноценная оптимизация с учетом специфики среды NVM
- Требуется тщательная и кропотливая работа для выбора подмножества стандарта, которое бы действительно требовалось в специализированной транзакционной СУБД

Заключение

- Доступная в ближайшем будущем энергонезависимая основная память будет обладать скоростью, не меньше, чем DRAM, и стойкостью, сопоставимой со стойкостью HDD (или хотя бы SSD)
- Исследования архитектур СУБД с одноуровневой системой хранения становятся более чем актуальными
- Даже наиболее удачные опубликованные исследования выполнялись при наличии ряда ограничений, не позволяющих получить «чистую» архитектуру транзакционной in-NVM СУБД
- Создан набросок такой архитектуры, в котором выделены важные аспекты распределения памяти, организации хранения таблиц и индексов, управления транзакциями и оптимизации запросов
- Масса интересных задач


С Новым годом!
www.davno.ru