VMware сдвигает акцент в проекте Capitola на CXL

Обзор последних инициатив VMware в области построения решений уровневой памяти на базе протокола CXL.

Проект Capitola – ориентация на CXL

На VMworld 2021 компания VMware анонсировала Project Capitola [2], инициативу по использованию больших объемов памяти, цель которой программное объединение нескольких уровней памяти разных типов для обеспечения единой модели потребления, прозрачной для приложений. Это программно-ориентированное (или программно-определяемое) многоуровневое распределение памяти скрывает сложность управления дополнительным уровнем памяти в дополнение к DRAM, предоставляя единое унифицированное адресное пространство памяти.

В свете объявления Intel[1)] VMware сместит акцент Project Capitola на технологии памяти на основе CXL и не будет поддерживать Intel Optane PMem для этого варианта использования. CXL — это поддерживаемое в отрасли межсоединение с когерентным кэшированием, представляющее новаторскую технологию расширения памяти способами, которые могут обеспечить гораздо большую ценность для клиентов. VMware по-прежнему полностью привержена решению проблем клиентов и в будущем выпустит многоуровневое программное обеспечение с такими технологиями, как CXL, что позволит использовать уровни памяти на основе CXL. Эти усилия также в конечном итоге приведут к концепции большой памяти VMware для обеспечения пула памяти и дезагрегации памяти.

Многоуровневая память VMware

Многоуровневую память VMware начало развивать около двух лет назад (в 2020 г.). Основная идея, которую развивает VMware – сделать ее максимально прозрачной для пользователей и приложений. Т.е., например, максимально избавить пользователей от каких-либо настроек и дать возможность запуска “старых” нагрузок (например, 3-5-летней давности) на новой инфраструктуре с многоуровневой памятью. Основные ее особенности (рис. 1, [3]):

– более высокая утилизация ядер и памяти, бОльшая емкость;

– снижение совокупной стоимости владения;

– бОльшая полоса пропускания (bandwidth);

– обеспечение минимального снижения производительности (например, за счет упреждающего “умного” размещения страниц в быстрой памяти);

– прозрачность — единственный адрес энергозависимой памяти:

  • не требуется проводить каких-либо изменений для гостевой ОС или изменений приложений (возможность запуска любой операционной системы);

  • ESX самостоятельно обрабатывает размещение страниц;

– использование DRS и vMotion для снижения рисков (эвристики распределения по уровням передаются в DRS; например, DRS может самостоятельно принимать решение о размещении нагрузки на “правильном” хосте);

– обеспечивается справедливость рабочих нагрузок – стабильная производительность;

– нулевые изменения конфигурации, не требуются специальные настройки уровней;

– поддерживается мониторинг конкретного процессора (мониторы vMMR как на уровне виртуальных машин, так и на уровне хоста.

1) Intel недавно объявила о сворачивании бизнеса Intel Optane (сообщение Intel о доходах за второй квартал, https://www.intc.com/news-events/press-releases/detail/1563/intel-reports-second-quarter-2022-financial-results). Они также взяли на себя обязательство поддерживать клиентов Optane для существующих линеек энергонезависимой памяти Optane и твердотельных накопителей Optane до конца срока службы. Условия гарантии на продукцию Intel остаются неизменными: обычная 5-летняя гарантия с даты продажи и техническая поддержка в течение гарантийного периода. Сюда входят серия PMem 100 («Apache Pass»), серия PMem 200 («Barlow Pass») и твердотельный накопитель DC P4800X («Cold Stream»), а также DC P5800X («Alder Stream»). Intel также продолжит разработку энергонезависимой памяти Intel Optane (PMem) серии 300 (кодовое название «Crow Pass») для масштабируемых процессоров Intel Xeon 4-го поколения (кодовое название «Sapphire Rapids»).
VMware планирует продолжить поддержку Intel Optane PMem с выпусками vSphere 7.x и 8.x во всех поддерживаемых в настоящее время конфигурациях. VMware также планирует продолжить поддержку твердотельных накопителей Intel Optane с выпусками vSAN 7.x и 8.x во всех поддерживаемых в настоящее время конфигурациях [1].

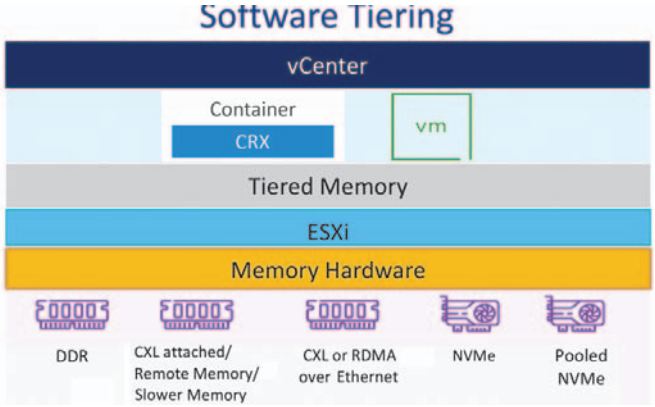

Рис. 1. Архитектура программно-определяемой уровневой памяти VMware.

VMware DRS используется для управления рабочими нагрузками для хостов VMware ESXi, разбитых на кластеры [5]. Основные функции, которые решает DRS:

– *балансировать вычислительную мощность* по кластерам:

  • повысить уровень обслуживания (SLA), гарантируя виртуальным машинам соответствующие ресурсы;

  • развертывать новую вычислительную мощность в кластере без прерывания обслуживания;

  • автоматически переносить виртуальные машины без прерывания обслуживания;

  • контролировать и управлять большим количеством инфраструктуры в расчете на одного системного администратора;

– *снизить энергопотребление* – динамически оптимизировать энергопотребление в кластере vSphere с помощью VMware vSphere Distributed Power Management (DPM), который также включен в редакции vSphere Enterprise Plus и vSphere with Operation Management Enterprise Plus. Когда потребность в ресурсах низкая, DPM переводит узлы в режим ожидания, а когда потребность высока, DPM включает достаточное количество узлов, чтобы справиться с этой потребностью и обеспечить доступность ваших служб. Динамическое управление питанием с помощью DPM позволяет:

  • сократить расходы на электроэнергию и охлаждение на целых 20% в периоды низкого использования;

  • автоматизировать управление энергопотреблением в центре обработки данных более эффективно;

– *обеспечить начальное размещение рабочей нагрузки* – когда включается виртуальная машина в кластере, DRS размещает ее на соответствующем хосте или создает рекомендацию, в зависимости от выбранного вами уровня автоматизации. Уровни автоматизации, также известные как пороги миграции, варьируются от консервативного до агрессивного. VMware vCenter будет применять только те рекомендации, которые удовлетворяют ограничениям кластера, таким как правила привязки хостов или обслуживание. Он применяет рекомендации DRS, которые могут обеспечить даже незначительное улучшение общего баланса нагрузки кластера. DRS предлагает пять уровней автоматизации в соответствии с вашими потребностями для каждого кластера;

– *автоматически балансировать нагрузку* – DRS распределяет рабочие нагрузки виртуальных машин между узлами vSphere внутри кластера и отслеживает доступные ресурсы. В зависимости от уровня автоматизации DRS перенесет (с помощью VMware vSphere vMotion) виртуальные машины на другие хосты в кластере, чтобы максимизировать производительность;

– *оптимизировать энергопотребление* – как и DRS, функция распределенного управления питанием (DPM) vSphere оптимизирует энергопотребление на уровне кластера и хоста. Когда запускается DPM, он сравнивает емкость на уровне кластера и узла с потребностью виртуальной машины, включая последние исторические потребности, и переводит узлы в режим ожида-

ния. Если потребности в емкости увеличиваются, DPM включает узлы в режиме ожидания, чтобы поглотить дополнительную рабочую нагрузку. Также можно настроить DPM так, чтобы он выдавал рекомендации, но не предпринимал никаких действий;

– *обеспечить обслуживание кластера* – DRS ускоряет процесс апдейта VMware vSphere Update Manager, определяя оптимальное количество хостов, которые могут одновременно переходить в режим обслуживания, исходя из текущих условий и потребностей кластера;

– *корректировать ограничения* – DRS перераспределяет виртуальные машины между хостами кластера vSphere для соответствия определяемым пользователем правилам привязки и антипривязки после сбоев хоста или во время операций обслуживания.

Функция vSphere Memory Monitoring and Remediation (vMMR), которая появилась в vSphere 7.0 U3. vMMR, помогает устранить потребность в мониторинге, предоставляя текущую статистику как на уровне виртуальной машины (пропускная способность), так и на уровне хоста (пропускная способность, частота промахов). vMMR также предоставляет оповещения по умолчанию и возможность настраивать пользовательские оповещения в зависимости от рабочих нагрузок, выполняемых на виртуальных машинах [4].

vMMR дает возможность администраторам отслеживать и настраивать использование и доступ к виртуальным машинам и приложениям, а также, при необходимости, выполнять исправления, перемещая виртуальные машины на хосты, чтобы они могли продолжать выполнять соглашения об уровне обслуживания. В будущих версиях vSphere на основе статистики, собранной vMMR, планировщик динамических ресурсов (DRS) сможет в определенных случаях выполнять автоматическое исправление .

Например, очень желательно, чтобы виртуальные машины на хосте имели справедливую долю использования DRAM с низкой задержкой для приложения «активный набор страниц памяти», в то же время сводя к минимуму использование постоянной памяти. Использование DRAM в качестве кэша иногда делает непредсказуемым использование приложениями DRAM по сравнению с энергонезависимой памятью, а по мере увеличения количества виртуальных машин и приложений обеспечение такой справедливости или гарантий SLA становится все более сложной задачей. vMMR предоставляет информацию об использовании памяти как для DRAM, так и для постоянной памяти, чтобы можно было предпринять соответствующие корректирующие действия для перемещения виртуальных машин на нужный хост .

Например, если виртуальная машина сталкивается с постоянным удалением памяти из-за промахов DRAM или если постоянное использование полосы пропускания памяти постоянно высокое, а производительность достигает критической стадии, как видно из статистики на основе vMMR, администраторы могут использовать VMware vMotion для переноса этой виртуальной машины на хост . который имеет больший объем DRAM и ресурсы постоянной памяти, доступные при просмотре статистики на целевом хосте. Это помогает оптимизировать операции и поддерживать стабильную производительность приложений.

Для обеспечения функции многоуровневой памяти VMware доработала ядро ESX, прежде всего, модули Host Memory Management

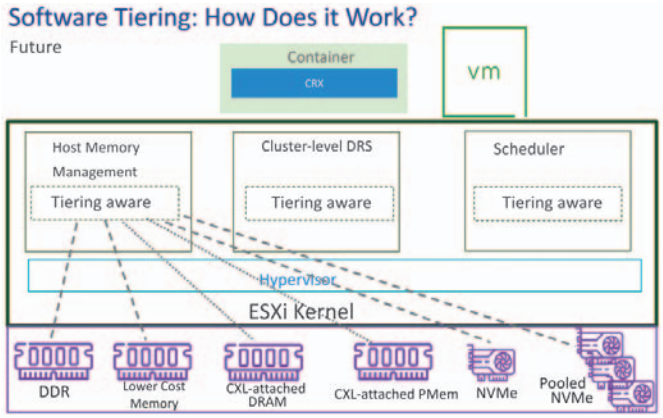

Рис. 2. В будущем число поддерживаемых уровней памяти будет расширено.

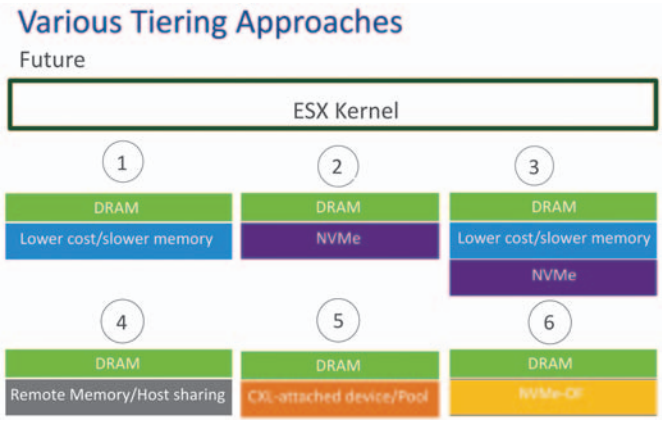

Рис. 3. Шесть возможных вариантов уровневой памяти.

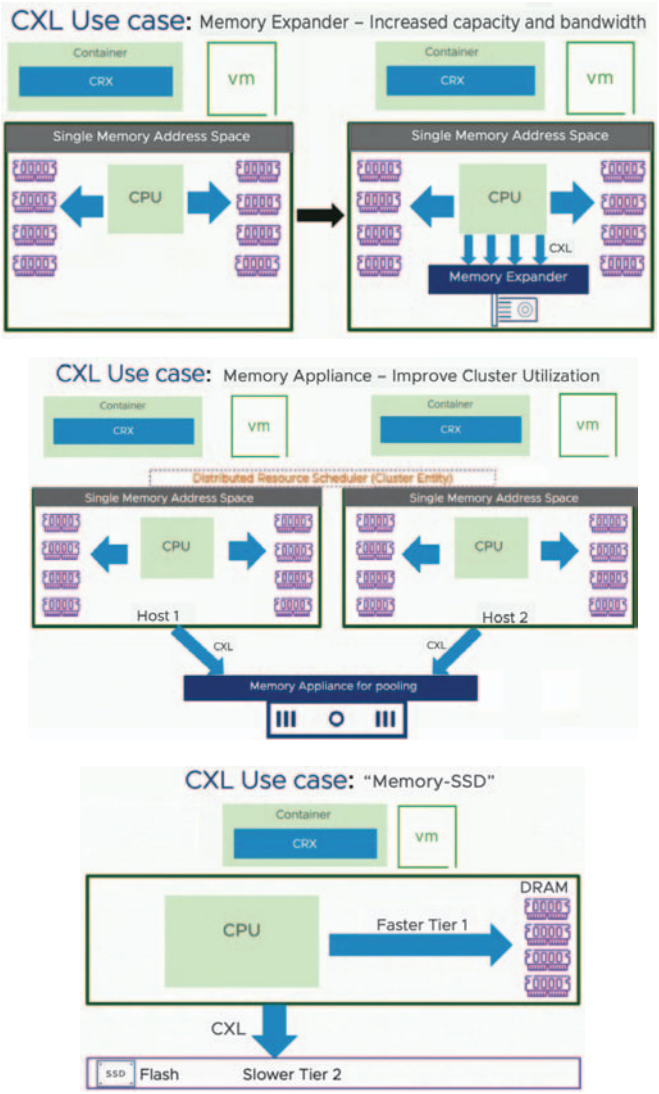

Рис. 4. Три варианта CXL-расширения памяти [6].

и DRS, добавив в них функцию Tiering aware. На первом этапе будет поддерживаться 2 уровня: DDR и Lower Cost Memory. В дальнейшем число возможных вариантов уровней пямяти будет расширено (рис. 2, 3). Рассматриваются шесть вариантов уровней памяти: DRAM + Lower cost/slower memory, DRAM + NVMe, DRAM + Lower cost/slower memory + NVMe, DRAM + Remote Memory/Host sharing, DRAM + CXL-attached device/Pool, DRAM + NVMe-OF. На первом этапе будет поддерживаться стандарт CXL 1.1.

VMware выделяет 3 варианта использования CXL (рис. 4) [6]:

– *memory expander* – увеличение емкости и полосы пропускания;

– *memory appliance* – улучшение утилизации кластера;

– *memory SSD.*

На VMware Explore 2022 [7,8] были представлены результаты тестирования уровневой памяти. Основные выводы и лучшие практики следующие:

– корпоративные приложения (реляционные базы данных, такие как SQL Server, и хранилища данных в памяти, такие как REDIS) могут извлечь выгоду из многоуровневой SW распределенной памяти:
  - в 2 раза больше плотности виртуальных машин в HammerDB, SQL Server (профиль OLTP/TPC-C);
  - в 2 раза больше плотности виртуальных машин в памяти REDIS;
  - на 30% выше производительность в HammerDB, SQL Server (профиль TPC-H/CLAP);
  - бесшовная миграция с помощью vMotion от системы DRAM к многоуровневой системе памяти;
– следите за своим монитором Active рабочего набора: если он меньше размера DRAM, хорошо подходит SW Memory Tiering;
– эксперты vSphere Performance будут вести видеоблоги о новых функциях и ключевых темах, включая, помимо прочего, распределение памяти по уровням: *https://blogs.vmware.com/performance/* – присоединяйтесь.

Источники, доп. ресурсы

[1] Intel Optane, Memory Optimization, and vSphere, August 28, 2022, Blog post by Paul Turner, Vice President, vSphere Product Management, VMware – *https://blogs.vmware.com/vsphere/2022/08/intel-optane-memory-optimization-and-vsphere.html*.

[2] Introducing Project Capitola: Software defined memory for data centric workloads, October 5, 2021, Blog post by Dave Morera and Ragav Gopalan, Cloud Platform Business Unit, VMware – *https://blogs.vmware.com/vsphere/2021/10/introducing-project-capitola.html*.

[3] Towards a CXL Future with VMware, Flash Memory Summit, August 4, 2022 – *https://www.youtube.com/watch?v=YxxFUOuTJaE&t=329s*.

[4] Persistent Memory technology as the new normal with VMware vSphere® Memory Monitoring and Remediation (vMMR), Arvind Jagannath, November 16, 2021 – *https://blogs.vmware.com/vsphere/2021/11/persistent-memory-technology-as-the-new-normal-with-vmware-vsphere-memory-monitoring-and-remediation-vmmr.html*.

[5] Distributed Resource Scheduler – *https://www.vmware.com/ru/products/vsphere/drs-dpm.html*.

[6] CXL Readiness with vSphere Virtuatization, Arvind Jagannath, Sr Product Line Manager, vSphere VMware, OCP Global Summit, October 18-20, 2022 – *https://www.youtube.com/watch?v=EIAwAeFvVw8*.

[7] Software Memory Tiering, VMware Explore 2022, Session code: CEIB1268USD – *https://www.vmware.com/explore/video-library.html#year=2022*.

[8] Extreme Performance Series 2022: Software-Based Transparent Memory Tiering Technology Prototype, August 30, 2022 – *https://www.youtube.com/watch?v=cJUf2weYzeI*.