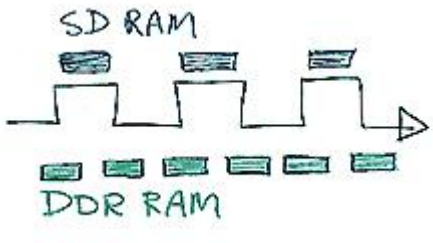
Оперативная память: поиск сложных решений

06 Фев 2016

Илья Манусов



По ряду объективных причин первым полигоном для испытания новых стандартов оперативной памяти становятся видео адаптеры. Судя по номенклатуре уже выпущенных устройств, технология вертикальной компоновки кристаллов и стандарт High Bandwidth Memory не станут исключением из такой закономерности. Было бы непростительным упущением выпустить из поля зрения столь важную инициативу.

Анатомия производительности

Оперативная память *High Bandwidth Memory (HBM)* для повышения пропускной способности использует шину данных увеличенной разрядности (от 1024 до 4096 бит). Для того чтобы такое решение стало возможным, как с технической так и экономической точки зрения, необходим метод компактного монтажа элементов памяти. Микросхемы размещаются в виде многоэтажной конструкции (*3D stacked*). Сигнальные линии пропускаются через такую многоэтажную конструкцию с использованием технологии *TSV (Through Silicon Via)*, в дословном переводе «*через кремниевые отверстия*».

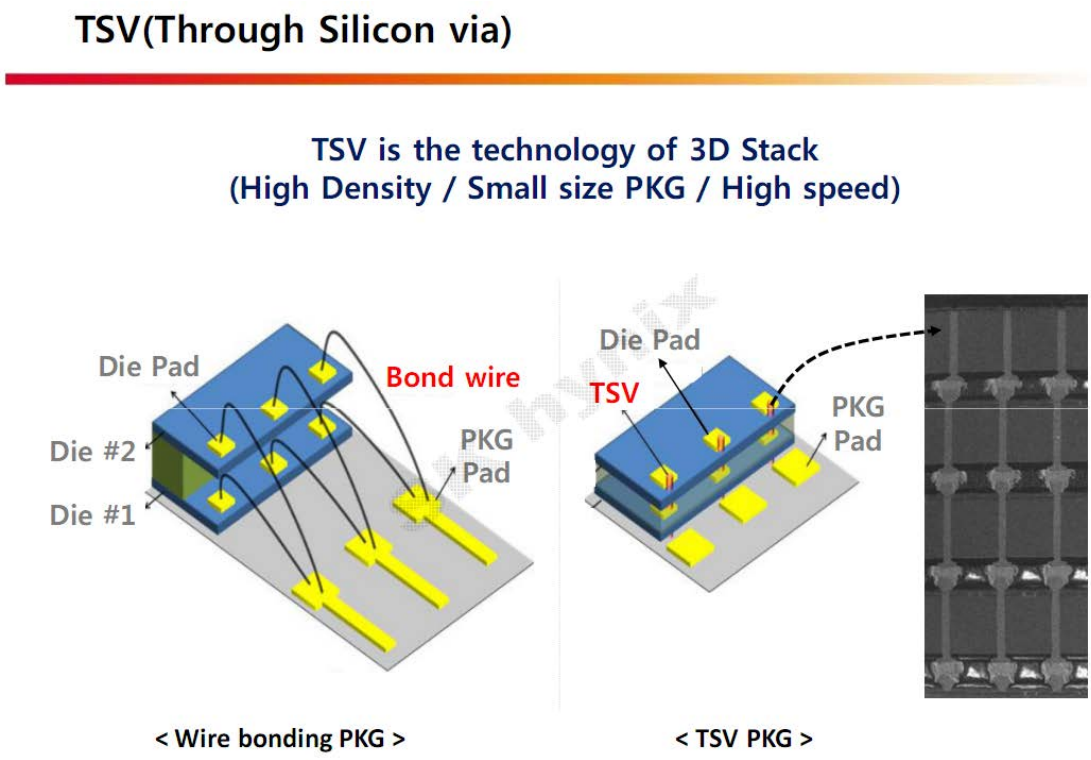


Рис 1. *Сравнение классического метода подключения и TSV. TSV позволяет пропустить сигнальные линии сквозь «многоэтажную конструкцию» из микросхем памяти, что существенно уменьшает занимаемую площадь и длину проводников*

Конечно, никто не запрещает экстенсивный подход — можно реализовать высокую разрядность просто поставив много микросхем на плату большой площади, пойдя на существенное увеличение стоимости и габаритов устройства, а также ограничив тактовую частоту, поскольку устойчивая синхронизация и передача сигналов в этом случае была бы затруднительной. В качестве примера одного из таких «устройств, опередивших время» можно вспомнить видеоадаптер Matrox Parhelia-512.

Итак, физический принцип, лежащий в основе HBM, предельно прост: увеличение разрядности шины данных. А «изюминка» новой памяти кроется в методе реализации. Контактные площадки размещаются на микросхеме памяти таким образом, что несколько микросхем можно установить на плате в виде многоэтажной конструкции, что существенно уменьшает занимаемую площадь и длину проводников, влияющую на устойчивость передачи сигналов высокой частоты.

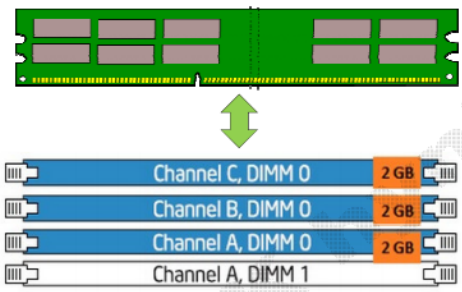
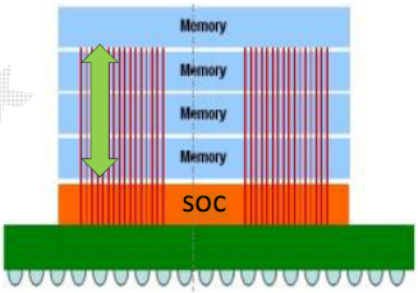
	DDR3	TSV(HBM)
Config.		
IO	64 DQ	1024 DQ
Speed	1.6G bps	1~2Gbps
Bandwidth	64 Gbps ➔ 12.8GBps	1024 Gbps ➔ Max 256GBps
Compare	Long Length➔ RLC increase	Short Length➔ RLC decrease

Рис 2. *Сравнение характеристик стандартного DDR3 DIMM и примера реализации подсистемы памяти на основе HBM. Разрядность шины данных удалось увеличить от 64 бит до 1024 бит без увеличения габаритов конструкции. Пропускная способность подсистемы памяти для операций чтения и записи, равная произведению пропускной способности одной сигнальной линии на количество сигнальных линий, увеличилась от 12.8 (PC3-12800) до 256 GBPS. Паразитные параметры, характеризующие электрофизическую устойчивость передачи сигналов, минимизированы в результате минимизации габаритов подсистемы памяти. Здесь RLC расшифровывается как: R = активное сопротивление линии связи L = индуктивность линии связи C = паразитная емкость*

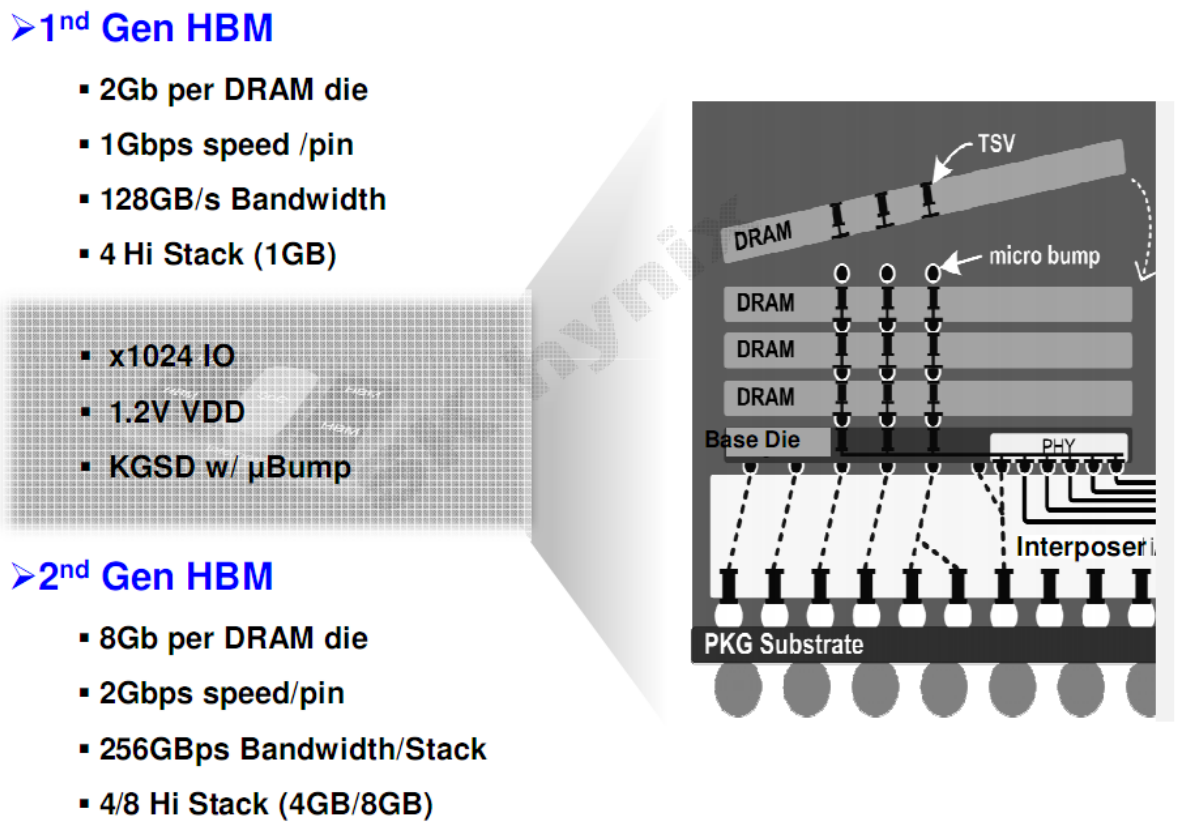


Рис 3. *Сравнение характеристик первого и второго поколений памяти HBM и схематическая иллюстрация метода монтажа с применением TSV. Критерии сравнения: емкость кристалла, пропускная способность в пересчете на одну линию, полная пропускная способность шины и количество «этажей» монтажа кристаллов памяти. Проблема отвода тепла в такой конструкции заслуживает быть темой отдельной статьи*

Разрядность — не панацея

Решения, основанные на повышении разрядности, эффективны для *потоковых* операций, при которых выполняется чтение или запись непрерывных блоков сравнительно большого размера. В первую очередь это актуально для видео адаптеров. При работе графического акселератора, а также при передаче отображаемых данных из фрейм-буфера в канал отображения, имеет место последовательный доступ, для которого такой подход эффективен. Антагонистический пример — при обработке фрагментированных, «мелко разбросанных» данных, когда необходимо прочитать несколько байтов, произвольно расположенных в адресном пространстве на большом расстоянии, шина данных высокой разрядности будет использоваться нерационально, так как в каждом 1024 или 4096-битном шинном цикле будет задействовано всего только 8 бит.

Вспомним и нашумевшую в свое время технологию RAMBUS, которая использует повышение частоты шины как альтернативу увеличению количества линий данных. Латентность (время реакции на изменившийся адрес) у нее также высокая, так как изменен подход к передаче сигналов и адресации банков, а не быстродействие самой ячейки памяти. Решение проблемы латентности и обеспечения производительности при работе с не потоковыми данными состояло бы в увеличении тактовой частоты самих ячеек DRAM, а не количества проводников и частоты передающих сигналы шин, но это значительно сложнее и затрагивает фундаментальные физические принципы, поэтому как HBM, так и RAMBUS делают упор на потоковые операции.

Резюме

Увеличение разрядности памяти, например, в видео адаптере, где это в первую очередь актуально, принесет пользу только тогда, когда остальные участники событий (CPU, GPU, видеодрайвер etc.) способны обрабатывать данные в увеличившемся темпе. Иначе, процесс будет напоминать соревнование между производственными линиями, выпускающими правый и левый ботинок. В любом случае, не следует забывать и о маркетинговой составляющей происходящих процессов...

Теги: Технологии