Сбалансированные конфигурации памяти процессоров семейства

AMD EPYC

Структура процессора AMD EPYC

Объяснение технологии чередования и ее важности

Контроллер памяти процессора

Советы по балансировке памяти

Подготовлено SMB-Solution

Обзор

Конфигурирование сервера со сбалансированной памятью важно для достижения максимальной пропускной способности подсистемы оперативной памяти и, как результат - общей производительности системы.

Содержание

Введение

Подсистема памяти является ключевым компонентом архитектуры всех процессоров и может существенно повлиять на его производительность. При правильном проектировании подсистема памяти может обеспечить чрезвычайно высокую пропускную способность и низкую задержку доступа к памяти. Если подсистема памяти неправильно спроектирована или модули памяти установлены в систему некорректно, общая производительность сервера может быть значительно снижена.

В этом кратком руководстве описывается подсистема памяти процессоров AMD EPYC и объясняется концепция сбалансированных конфигураций памяти, которые обеспечивают

максимально возможную пропускную способность.

В обзоре рассматриваются процессоры семейства AMD EPYC. Семейства процессоров Intel Xeon Scalable, Intel E5 v4 и Intel E7 v4 обсуждались в предыдущих кратких обзорах:
«Сбалансированные конфигурации памяти процессоров семейства Intel Xeon Scalable»
«Достижение максимальной производительности System x и ThinkServer с использованием балансировки конфигурации памяти».

Структура процессора AMD EPYC

AMD EPYC строятся по технологии MCM (Multi Chip Module) и состоят из 4 процессорных модулей (кристаллов) архитектуры Zeppelin, расположенных на одной подложке и объединенных в единое целое шинами Infinity Fabric.

Такое решение позволяет повысить выход годных изделий, поскольку каждый из процессорных модулей Zeppelin имеет площадь 213 мм² (суммарная площадь процессора из 4 модулей — 852 мм², что по сегодняшним меркам слишком велико для используемого фотолитографического процесса).

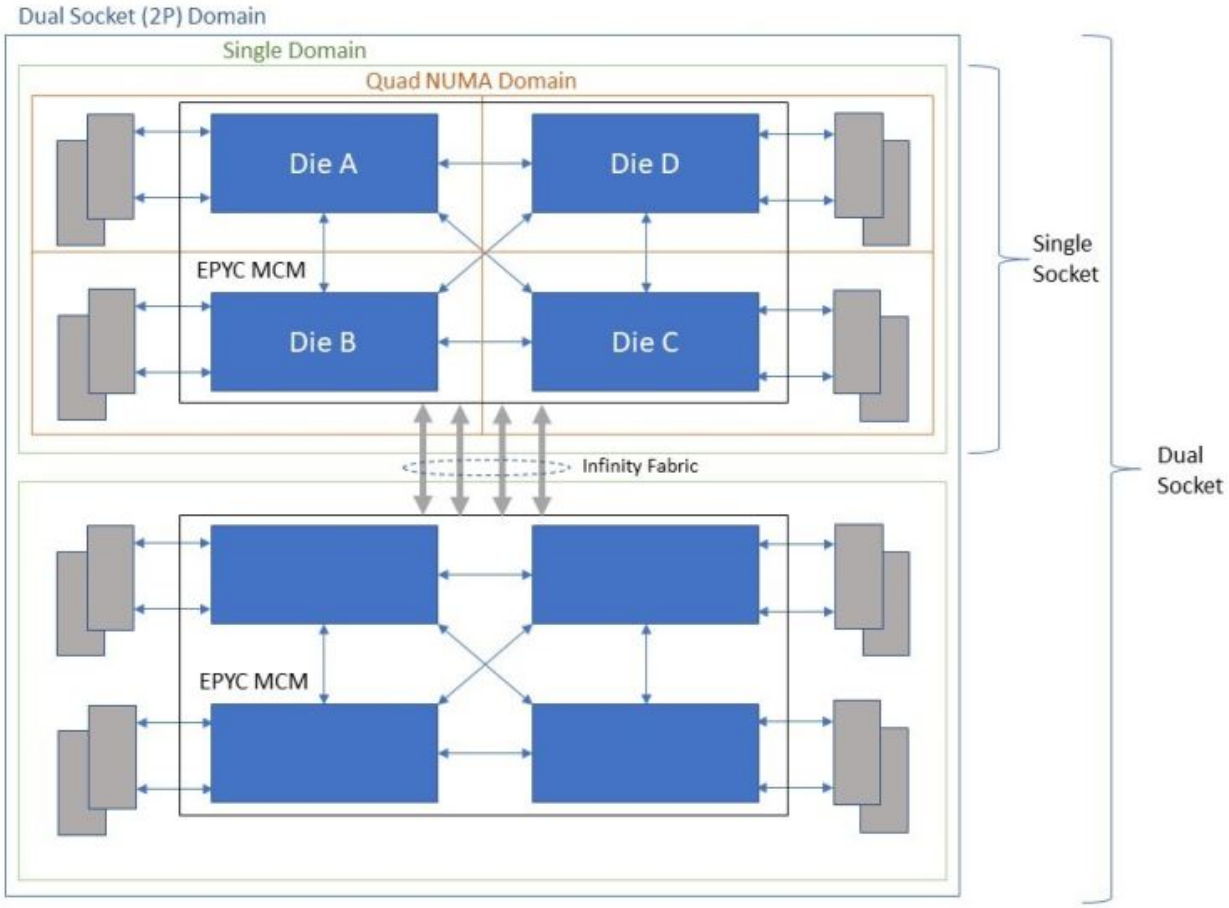


Рис. 1 Структурная схема процессора AMD EPYC

Каждый процессорный модуль содержит от 2 до 8 процессорных ядер (после отбраковки ничего не выбрасываем, все идет в производство!), 32 линии PCIe и двухканальный контроллер памяти. В сумме на один процессор EPYC приходится от 8 (2 ядра на модуль) до 32 (8 ядер на модуль) ядер в зависимости от модели. При этом, абсолютно все процессоры EPYC имеют 128 (4x32) линий PCIe и 8 каналов памяти. Несколько моделей предназначены исключительно для однопроцессорных

серверов (индекс P на конце, например: 7401P, 7351P), остальные могут устанавливаться как в одно- , так и в двухпроцессорные системы.

Таблица параметров процессоров приведена ниже:

Модель	Ядер	Потоков	Базовая частота, MHz	Максимальная частота (все ядра), MHz	Максимальная частота, MHz	Потребляемая мощность, Вт	Кэш L3, Mb	Каналов памяти	Максимальная частота памяти (1 модуль на канал)	Теоретическая пропускная способность 2-процессорной системы памяти, GB/s	Линий PCIe	Одно / двухпроцессорные системы	Применение
7601	32	64	2.20	2.70	3.20	180	64	8	2666	341	x128	2P/1P	СУБД, Аналитика, HPC
7551	32	64	2.00	2.55	3.00	180	64	8	2666	341	x128	2P/1P	VM, VDI, СУБД, Аналитика, HPC
7551P												1P	
7501	32	64	2.00	2.60	3.00	155/170	64	8	2400/2666	307/341	x128	2P/1P	VM, VDI, СУБД, Аналитика, WEB сервисы
7451	24	48	2.30	2.90	3.20	180	64	8	2666	341	x128	2P/1P	общего назначения
7401	24	48	2.00	2.80	3.00	155/170	64	8	2400/2666	307/341	x128	2P/1P	общего назначения,GPU / FPGA ускорители, хранилища
7401P												1P	
7351	16	32	2.40	2.90	2.90	155/170	64	8	2400/2666	307/341	x128	2P/1P	общего назначения,GPU / FPGA ускорители, хранилища
7351P												1P	
7301	16	32	2.20	2.70	2.70	155/170	64	8	2400/2666	307/341	x128	2P/1P	общего назначения
7281	16	32	2.10	2.70	2.70	155/170	64	8	2400/2666	307/341	x128	2P/1P	общего назначения
7251	8	16	2.10	2.90	2.90	120	64	8	2400	307	x128	2P/1P	общего назначения

Контроллер памяти процессора AMD EPYC

Доступ к оперативной информации, хранящейся на модулях DIMM, обеспечивается контроллерами памяти, которые интегрированы в процессор.

Как видно из схемы на Рис. 1, каждый из модулей процессора Zeppelin имеет собственный двухканальный контроллер памяти. Каждый из каналов позволяет установить до двух модулей DIMM. Итого, максимальное количество модулей DIMM, которое можно установить в систему:
- для 1-процессорной системы: 2 модуля на канал X 2 канала X 4 модуля на процессор = 16 модулей;
- для 2-процессорной системы: 2 модуля на канал X 2 канала X 4 модуля на процессор X 2 процессора = 32 модуля.

Процессоры AMD EPYC, так же, как и Intel Xeon, используют технологию неравномерного доступа к памяти (NUMA).

> Что такое NUMA? Многопроцессорная (или многомодульная) архитектура, в которой каждый процессор подключен к своей собственной локальной памяти (называемой доменом NUMA), но также может обращаться к памяти, подключенной к другому процессору.
>
> Технология называется «неравномерной», поскольку доступ к локальной памяти имеет более низкую задержку (память в своем NUMA-домене), чем когда требуется доступ к памяти, подключенной к домену NUMA другого процессора. Эта особенность учитывается в рекомендациях по установке модулей DDR в систему.

> Преимущество архитектуры NUMA заключается в том, что она обеспечивает многопроцессорную масштабируемость, увеличивает пропускную способность памяти с добавлением большего количества процессоров и уменьшает конфликт памяти для процессоров, если они конкурируют за доступ через общую шину.
>
> NUMA в реализации AMD уменьшает задержки памяти и уменьшает трафик передачи данных, по сравнению с решением Intel Manhattan Mesh за счет дополнительных диагональных соединений между кристаллами.
>
> В идеальной среде выполнения программного обеспечения каждый процессорный сокет имеет достаточно памяти для выполнения всех потоков. Но это не так для многих современных рабочих нагрузок. Часто объем данных слишком велик, чтобы содержаться в памяти, установленной в одном сокете (подключенной к одному процессору). Иногда рабочие нагрузки конкурируют за локальные ресурсы и создают локальные узкие места. При использовании конфигурации с двумя сокетами (процессорами), задержка доступа к памяти будет зачительно увеличена, если трафик проходит через межсоединение процессоров. Происходит это независимо от того, является ли это соединение AMD Infinity Fabric или Intel QPI. В двухъядерных конструкциях от AMD или Intel планировщик NUMA должен размещать потоки и данные на ядрах в одном и том же сокете, чтобы уменьшить задержки. В противном случае, запросы данных между сокетами приводят к более высоким задержкам.

Контроллер памяти AMD EPYC может работать по всем каналам с частотой до 2400/2666 МГц. При этом, необходимо помнить, что установка двух модулей на один канал приводит к дополнительной нагрузке на шину памяти и — соответственно, к снижению частоты шины.

Общая пропускная способность подсистемы оперативной памяти процессоров AMD EPYC, в теории, имеет преимущество по сравнению с процессорами Intel Xeon.

Для двухпроцессорных систем EPYC:

2,4 GT/s x 8 байт на канал x 8 каналов x 2 сокета = 307 GB/s для 8 каналов на частоте 2400

2,66 GT/s x 8 байт на канал x 8 каналов x 2 сокета = 341 GB/s для 8 каналов на частоте 2666

Для двухпроцессорных систем Xeon:

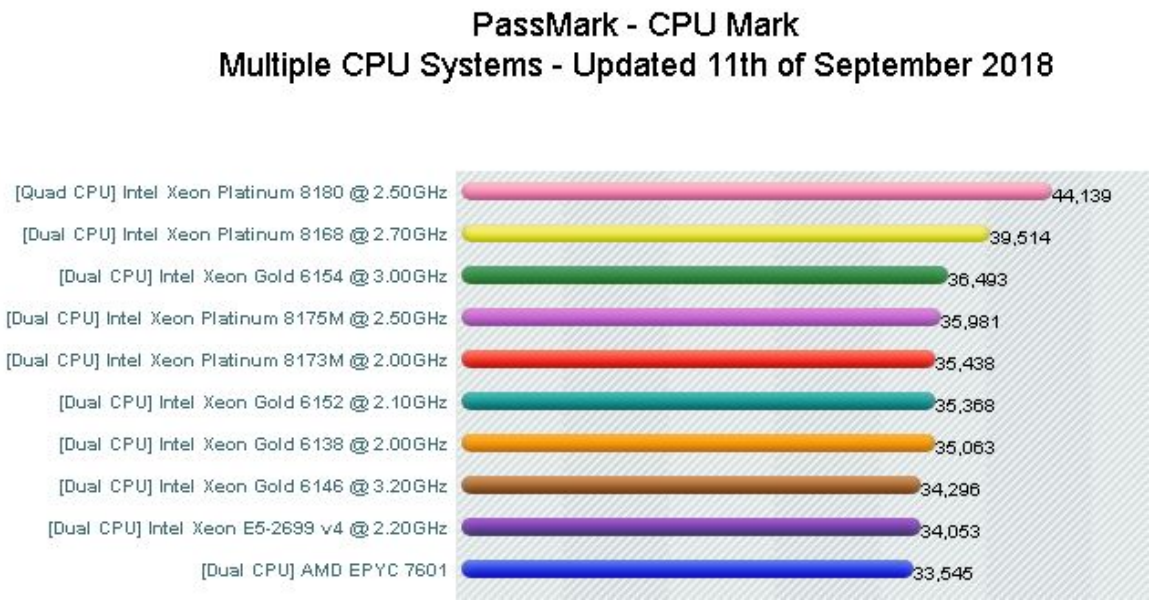
2,66 GT/s x 8 байт на канал x 6 каналов x 2 сокета = 255 GB/s для 6 каналов на частоте 2666

Понятно, что ограничение пропускной способности может проявиться в системах с высокой вычислительной нагрузкой и большим объемом данных, обрабатываемых в оперативной памяти. Задачи, связанные с большой долей дисковых операций, не заметят ограничения пропускной способности — задержки обращений к дисковой системе будут играть в них решающую роль.

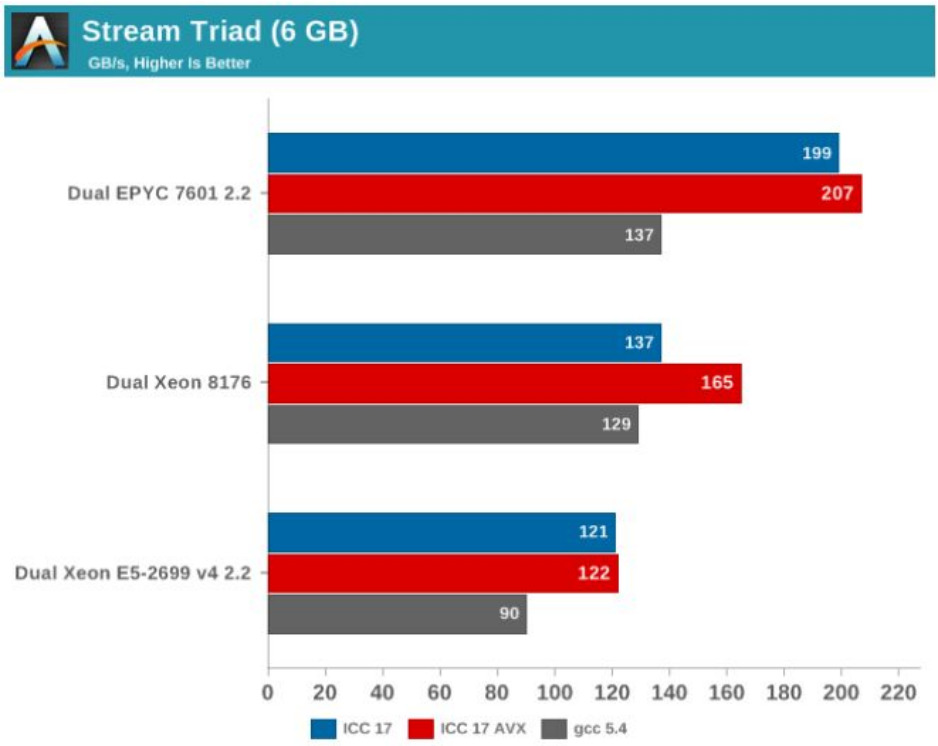
Синтетические тесты производительности систем

Теория — теорией, а вот что показывает практика: мы приводим сравнительные результаты тестирования из двух источников — PassMark и Anandtech.

PassMark использует свой набор синтетических тестов.



Anandtech проводил тестирование с помощью Stream 5.10, который также использовали авторы обзора компании Lenovo для процессоров Intel Xeon. Компания Anandtech произвела тестирование тремя вариантами теста — скомпилированных Intel compiler (ICC), Intel compiler (ICC) с векторными инструкциями AVX и GCC 5.4 compiler.



Source: Anandtech, July, 2017

Как видим - «мнения разошлись». У нас больше доверия вызывают результаты Anandtech на основе тестов Stream как более объективный, поскольку именно этот тест выбрали, как уже

указывалось, инженеры Lenovo для оценки пропускной способности памяти процессоров Intel Xeon.

Балансировка памяти

Современные процессоры, как Intel, так и AMD, имеют механизмы чередования обращений к памяти для уменьшения влияния задержек. Убрать задержки невозможно, но «замаскировать» их частично за счет распараллеливания операций удается.

Вместо последовательного обращения к ряду адресов одного устройства, чередование организует смежные области памяти на различных устройствах и обращается к ним поочередно. Следующее обращение производится еще до получения ответа на текущее — тем самым, время ожидания сокращается за счет совмещения операций во времени.

Чередование может быть организовано между:

- ранками в пределах одного модуля DIMM или нескольких DIMM;
- каналами контроллера;
- контроллерами процессора
- процессорами.

Еще один уровень чередования может быть организован ресурсами BIOS. Используются возможности объединения ресурсов в узел (node) и NUMA – когда части процессора жестко привязывают набор модулей DIMM как «хозяйский». Фактически, средствами BIOS один физический процессор разделяется на два независимых виртуальных, каждому из которых назначается набор модулей DIMM в единоличное пользование. Этот тип чередования может быть реализован, если никак не удается добиться симметрии, общее количество модулей DIMM не кратно количеству каналов памяти или используются модули разной емкости.

Проектирование сервера без учета балансировки памяти, так же — как установка модулей DIMM без учета балансировки, может привести к тому, что подсистема памяти будет работать неоптимально, с лишними задержками и использовать лишь 17% от потенциальной пропускной способности — эти расчеты приводятся в руководстве инженеров Lenovo для процессоров Intel. Для серверов с высокой вычислительной нагрузкой несоблюдение балансировки может иметь катастрофические последствия и вызвать неоправданные дополнительные вложения средств.

Сбалансированные конфигурации памяти

Основные принципы сбалансированной подсистемы памяти заключаются в следующем:

1. Все заполненные каналы памяти должны иметь одинаковую общую емкость модулей памяти и одинаковое общее количество ранков.
2. Все контроллеры памяти в процессорном сокете должны иметь одинаковую конфигурацию модулей DIMM.
3. Все сокеты процессора на одном физическом сервере должны иметь одинаковую конфигурацию модулей DIMM.

В общем, чем больше симметрии — тем лучше для производительности системы. Не забываем также, что количество установленных модулей DIMM влияет на возможности чередования - чем больше контроллеров и каналов памяти задействуется (в пределах одного модуля на канал, 1DPC) - тем система производительнее. К примеру - в потенциале, установка 16шт модулей емкостью 8Гб с

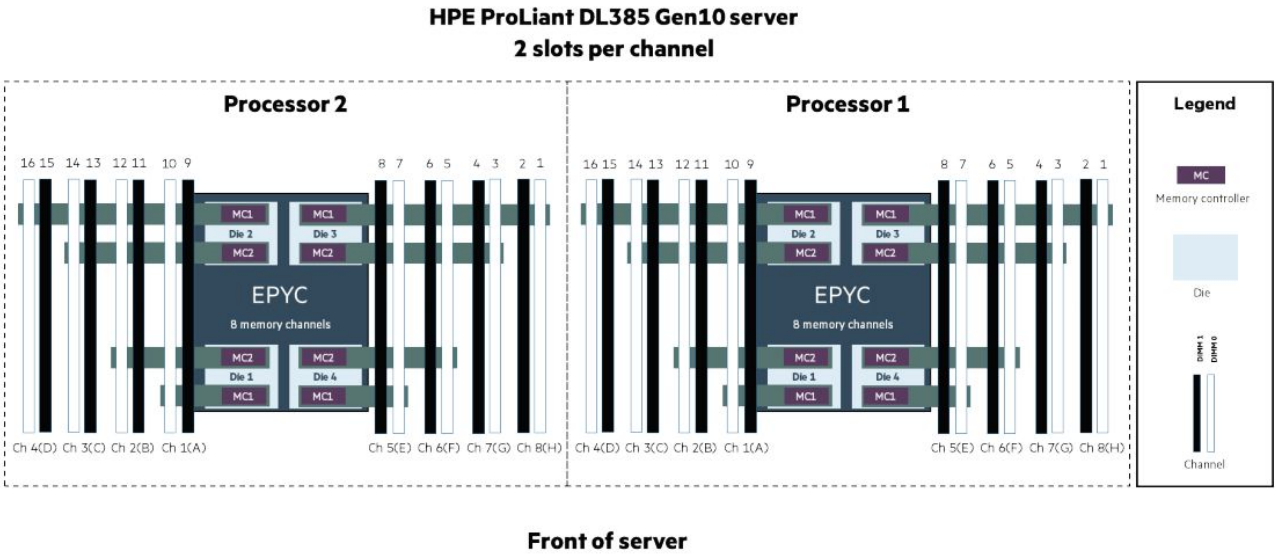
точки зрения производительности предпочтительнее, чем тот же объем оперативной памяти, но набранный меньшим количеством более емких модулей — 8шт по 16Гб или 4шт по 32Гб. Как показывает опыт - «задел» на будущее расширение оперативной памяти не работает. Или система никогда не расширяется (в подавляющем большинстве случаев), или - к моменту принятия решения о расширении, стоимость замены всех модулей памяти на более емкие оказывается ниже, чем добавление модулей.

Для достижения оптимальной пропускной способности памяти выполните следующие действия:
1. Оцените объем необходимой в сервере памяти,
2. Разделите этот резмер памяти на шестнадцать (для двухпроцессорной системы), чтобы определить минимальную требуемую емкость модуля DIMM,
3. Округлите эту рассчитанную емкость DIMM до ближайшей доступной емкости DIMM и
4. Установите в ваш сервер шестнадцать DIMM этой емкости
Например, если требуется 200 ГБ оперативной памяти, каждый модуль DIMM должен быть чуть более 12,5 ГБ. Ближайший размер доступной памяти DIMM составляет 16 ГБ. Суммарный объем, в таком случае, получится 16 x 16 = 256 Гб.

Примеры сбалансированных конфигурации памяти

Для иллюстрации, приведем примеры сбалансированных конфигураций для серверов HPE и Supermicro.

Двухпроцессорный сервер HPE ProLiant DL385 Gen10 имеет 32 слота для установки DIMM.



В первую очередь, заполняются белые слоты — первый модуль на канале. Только при нехватке количества слотов, заполняют цветные слоты.

Последовательность заполнения приведена в следующей таблице:

HPE ProLiant Gen10 servers—AMD two-processor configuration
DIMM population order

Number of DIMM(s) to populate	Processor 2																Processor 1													
	CH D		CH C		CH B		CH A		CH E		CH F		CH G		CH H		CH D		CH C		CH B		CH A		CH E		CH F		CH G		CH H	
1																	16														
2	16																16														
3	16																16														1
4	16															1	16														1
5	16															1	16				12										1
6	16				12											1	16				12										1
7	16				12											1	16				12						5				1
8	16				12							5				1	16				12						5				1
9	16				12							5				1	16		14		12						5				1
10	16		14		12							5				1	16		14		12						5				1
11	16		14		12							5				1	16		14		12						5		3		1
12	16		14		12							5		3		1	16		14		12						5		3		1
13	16		14		12							5		3		1	16		14		12		10				5		3		1
14	16		14		12		10					5		3		1	16		14		12		10				5		3		1
15	16		14		12		10					5		3		1	16		14		12		10			7	5		3		1
16	16		14		12		10			7		5		3		1	16		14		12		10			7	5		3		1
17	16		14		12		10			7		5		3		1	16	15	14		12		10			7	5		3		1
18	16	15	14		12		10			7		5		3		1	16	15	14		12		10			7	5		3		1
19	16	15	14		12		10			7		5		3		1	16	15	14		12		10			7	5		3	2	1
20	16	15	14		12		10			7		5		3	2	1	16	15	14		12		10			7	5		3	2	1
21	16	15	14		12		10			7		5		3	2	1	16	15	14		12	11	10			7	5		3	2	1
22	16	15	14		12	11	10			7		5		3	2	1	16	15	14		12	11	10			7	5		3	2	1
23	16	15	14		12	11	10			7		5		3	2	1	16	15	14		12	11	10			7	6	5		3	2	1
24	16	15	14		12	11	10			7	6	5		3	2	1	16	15	14		12	11	10			7	6	5		3	2	1
25	16	15	14		12	11	10			7	6	5		3	2	1	16	15	14	13	12	11	10			7	6	5		3	2	1
26	16	15	14	13	12	11	10			7	6	5		3	2	1	16	15	14	13	12	11	10			7	6	5		3	2	1
27	16	15	14	13	12	11	10			7	6	5		3	2	1	16	15	14	13	12	11	10			7	6	5	4	3	2	1
28	16	15	14	13	12	11	10			7	6	5	4	3	2	1	16	15	14	13	12	11	10			7	6	5	4	3	2	1
29	16	15	14	13	12	11	10			7	6	5	4	3	2	1	16	15	14	13	12	11	10	9		7	6	5	4	3	2	1
30	16	15	14	13	12	11	10	9		7	6	5	4	3	2	1	16	15	14	13	12	11	10	9		7	6	5	4	3	2	1
31	16	15	14	13	12	11	10	9		7	6	5	4	3	2	1	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
32	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

А теперь приведем таблицу оптимальных конфигураций памяти (для одного процессора!) для сервера HPE из руководства пользователя:

Gen10 AMD optimal memory configurations per processor

Memory (per processor)	Best configuration & resulting speed	Configuration
64 GB	8X 8 GB SR -> 2667 MTS	1 DPC
128 GB	8X 16 GB SR -> 2667 MTS	1 DPC
256 GB	8X 32 GB DR -> 2400 MTS	1 DPC
512 GB	8X 64 GB LR -> 2667 MTS	1 DPC
1024 GB	8X 128 GB LR -> 2667 MTS	1 DPC
2048 GB	16X 128 GB LR -> 2133 MTS	2 DPC

Пояснения: 1DPC – один модуль на канал (DIMM per Channel), 2DPC – два модуля на канал (DIMM per Channel). SR – одноранковые модули (Single Rank), DR – двухранковые модули (Double Rank). LR – Load Reduced модули.

В таблице также указана частота, на которй модули DIMM будут работать в приведенной конфигурации.

Следующая таблица показывает, как именно должны быть установлены модули DIMM при оптимальной конфигурации:

HPE ProLiant Gen10 AMD 2 Processor Configuration
DIMM population order

Number of DIMM(s) to populate	Processor 2																Processor 1													
	CH D		CH C		CH B		CH A		CH E		CH F		CH G		CH H		CH D		CH C		CH B		CH A		CH E		CH F		CH G		CH H	
1																1	16															
2	16																16															
3	16																16															1
4	16															1	16															1
5	16															1	16				12											1
6	16				12											1	16				12											1
7	16				12											1	16				12						5					1
8*	16				12						5					1	16				12						5					1
9	16				12						5					1	16		14		12						5					1
10	16		14		12						5					1	16		14		12						5					1
11	16		14		12						5					1	16		14		12						5		3			1
12	16		14		12						5		3			1	16		14		12						5		3			1
13	16		14		12						5		3			1	16		14		12		10				5		3			1
14	16		14		12		10				5		3			1	16		14		12		10				5		3			1
15	16		14		12		10				5		3			1	16		14		12		10			7	5		3			1
16	16		14		12		10		7		5		3			1	16		14		12		10			7	5		3			1
17	16		14		12		10		7		5		3			1	16	15	14		12		10			7	5		3			1
18	16	15	14		12		10		7		5		3			1	16	15	14		12		10			7	5		3			1
19	16	15	14		12		10		7		5		3			1	16	15	14		12		10			7	5		3		2	1
20	16	15	14		12		10		7		5		3		2	1	16	15	14		12		10			7	5		3		2	1
21	16	15	14		12		10		7		5		3		2	1	16	15	14		12	11	10			7	5		3		2	1
22	16	15	14		12	11	10		7		5		3		2	1	16	15	14		12	11	10			7	5		3		2	1
23	16	15	14		12	11	10		7		5		3		2	1	16	15	14		12	11	10			7	6	5		3	2	1
24	16	15	14		12	11	10		7	6	5		3		2	1	16	15	14		12	11	10			7	6	5		3	2	1
25	16	15	14		12	11	10		7	6	5		3		2	1	16	15	14	13	12	11	10			7	6	5		3	2	1
26	16	15	14	13	12	11	10		7	6	5		3		2	1	16	15	14	13	12	11	10			7	6	5		3	2	1
27	16	15	14	13	12	11	10		7	6	5		3		2	1	16	15	14	13	12	11	10			7	6	5	4	3	2	1
28	16	15	14	13	12	11	10		7	6	5	4	3	2	1		16	15	14	13	12	11	10			7	6	5	4	3	2	1
29	16	15	14	13	12	11	10		7	6	5	4	3	2	1		16	15	14	13	12	11	10	9		7	6	5	4	3	2	1
30	16	15	14	13	12	11	10	9	7	6	5	4	3	2	1		16	15	14	13	12	11	10	9		7	6	5	4	3	2	1
31	16	15	14	13	12	11	10	9	7	6	5	4	3	2	1		16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
32	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	16	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1

Balanced Configuration

*=Channel Interleaving not Supported

Пояснения: серым выделены строки сбалансированных конфигураций. Вариант с 8 модулями хоть и относится к сбалансированным, но не поддерживает чередование каналов.

Двухпроцессорные серверы Supermicro также имеют 32 слота для установки DIMM.

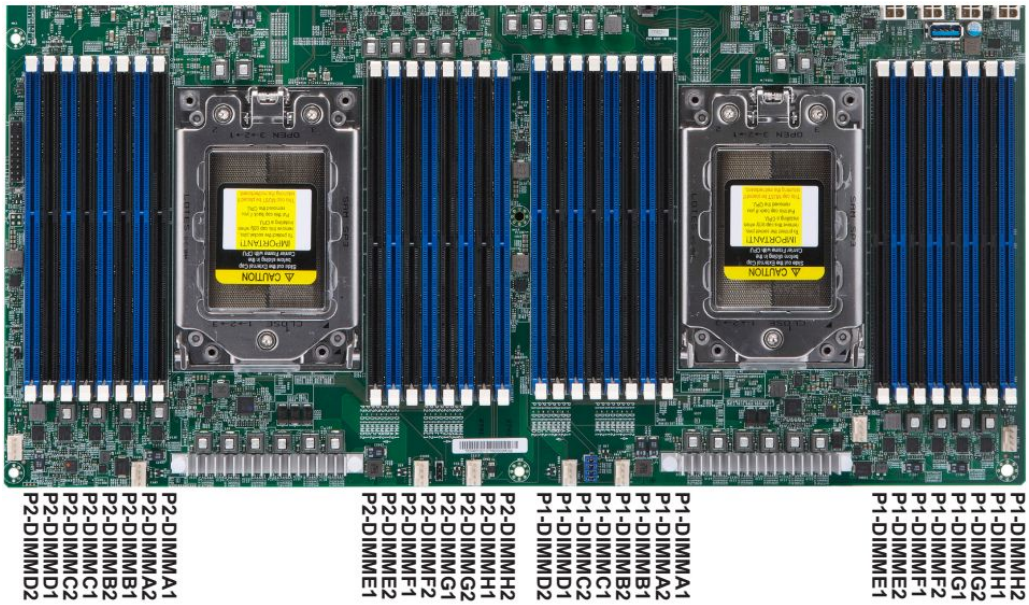


Figure 3-12. DIMM Numbering

Голубые слоты заполняются в первую очередь — это первый модуль на канале.

Последовательность заполнения модулей для одного процессора следующая:

DIMMA2, DIMMB2, DIMMC2, DIMMD2, DIMME2, DIMMF2, DIMMG2, DIMMH2,

затем:

DIMMA1, DIMMB1, DIMMC1, DIMMD1, DIMME1, DIMMF1, DIMMG1, DIMMH1.

То есть, заполнение ведем от процессора в одну сторону, затем — в другую. Только после заполнения голубых слотов, начинаем аналогично заполнять черные — вторым модулем DIMM на канал.

Следующая таблица показывает оптимальные конфигурации памяти для одного процессора в зависимости от емкости модуля DIMM и частоту, на которой память будет работать:

Populating RDIMM/RDIMM 3DS/LRDIMM/LRDIMM 3DS DDR4 Memory Modules

Type	DIMM Population		Maximum DIMM Capacity (GB)		Maximum Frequency (MHz)
	DIMM1	DIMM2	1 Channel	8 Channel	
RDIMM		1R	16GB	128GB	2666
	1R	1R	32GB	256GB	2133
		2R	32GB	256GB	2400
	1R	2R	48GB	384GB	1866
	2R	2R	64GB	512GB	1866
LRDIMM		4R	64GB	512GB	2666
	4R	4R	128GB	1TB	2133
		8R	128GB	1TB	2666
	4R	8R	192GB	1.5TB	2133
	8R	8R	256GB	2TB	2133
LRDIMM 3DS		2R2H	64GB	512GB	2400
	2R2H	2R2H	128GB	1TB	1866
		2R4H	128GB	1TB	2400
	2R2H	2R4H	192GB	1.5TB	1866
	2R4H	2R4H	256GB	2TB	1866

Источники:

https://www.amd.com/system/Ffiles/2018-03/AMD-Optimizes-EPYC-Memory-With-NUMA.pdf

NUMA, Non-Uniform Memory Access: https://ru.wikipedia.org/wiki/Non-Uniform_Memory_Access

Supermicro AS-1123US-TR4 руководство пользователя:
https://supermicro.com/manuals/superserver/1U/MNL-1960.pdf

Server memory population rules for HPE ProLiant Gen10 servers:
https://h20195.www2.hpe.com/v2/GetDocument.aspx?docname=a00038346enw