TRUE SYSTEM
STORAGE SOLUTION COMPANY

Intel Xeon E5-1600/2600 v3/v4 и оперативная память

Дмитрий Носачёв, nosachevd@truesystem.ru

24 марта 2017 г.

Аннотация

В данной статье рассматриваются вопросы работы оперативной памяти с платформой Intel Grantley, состоящей из чипсета Intel C612 в связке с процессорами Intel Xeon E5-2600 v3 (Haswell-EP) и E5-2600 v4 (Broadwell-EP).

Правила установки модулей памяти, нумерация разъёмов, сведения о настройке режимов работы памяти в BIOS указаны на примере материнской платы Supermicro X10DRi-LN4+. Информация из этой статьи применима к другим продуктам производства Supermicro или других компаний на базе чипсета Intel C612, но за уточнениями мы советуем обращаться к документации производителя.

1 Компоненты

Как и процессоры предыдущих поколений[1], Haswell-EP и Broadwell-EP имеют встроенный 4-канальный контроллер оперативной памяти[2], то есть каждый процессор непосредственно контролирует свой набор модулей памяти. Обращение к памяти соседнего процессора осуществляется через межпроцессорную шину QPI. Подобная архитектура получила название NUMA[3].

1.1 Сравнительные характеристики процессоров E5-1600/2600 v4

Основные характеристики процессоров Intel Xeon E5-2600 и E5-1600 v4/v3 (количество ядер, тепловой пакет, тактовая частота в обычном режиме и режиме Turbo Boost, максимальная частота работы памяти) представлены в таблицах 1, 2, 4, 3. Следует помнить, что процессоры E5-1600 не предназначены для работы в 2-процессорных платах.

[1] Intel Xeon E5-2600 (Sandy Bridge-EP) и Intel Xeon E5-2600 v2 (Ivy Bridge-EP).

[2] В старших моделях используется два 2-канальных контроллера с целью увеличения производительности. Также эта особенность используется в технологии Cluster on Die.

[3] Non-Uniform Memory Access — «неравномерный доступ к памяти» или Non-Uniform Memory Architecture — «Архитектура с неравномерной памятью».

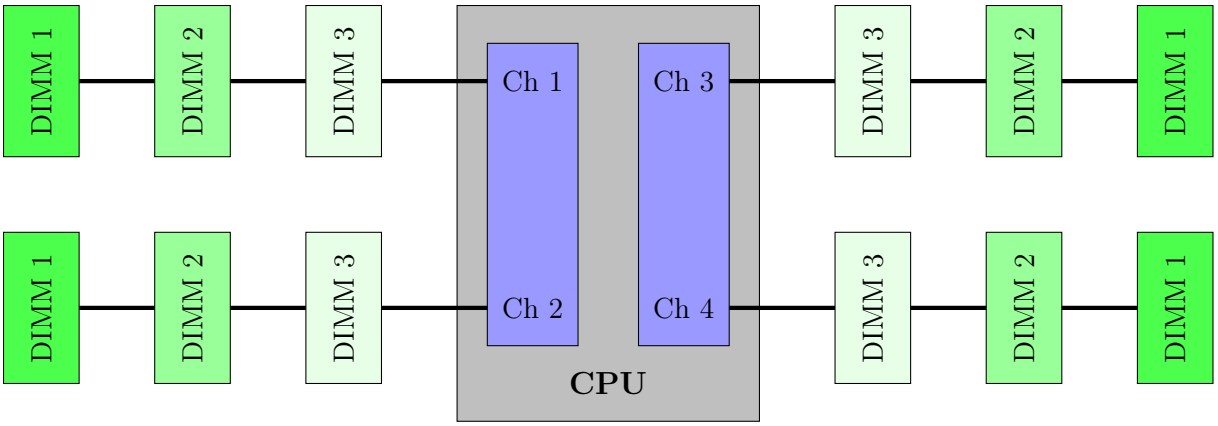

Рис. 1: Схема подключения оперативной памяти к процессору Intel Xeon E5.

1.2 Материнские платы

Двухпроцессорные материнские платы производства Supermicro на базе чипсета Intel C612 могут иметь 8/16/24 разъёмов для установки модулей памяти (4, 8 или 12 модулей на процессор). Схема установки модулей памяти при наличии 12 разъёмов DIMM на каждый процессор (3 на канал) показана на рисунке 1. Некоторые серверы Supermicro требуют установки низкопрофильных модулей памяти (VLP, Very Low Profile).

На материнских платах Supermicro разъёмы DIMM промаркированы номером процессора, буквой латинского алфавита, обозначающей канал (A–D — 4 канала первого процессора, E–H — 4 канала второго процессора,), и порядковым номером установки модуля. Примеры:

- **P1-DIMMA2** — первый процессор, первый канал (A), разъём 2.
- **P2-DIMMG1** — второй процессор, третий канал (G), разъём 1.

Первые разъёмы DIMM каждого канала имеют синий цвет (см. рисунок 2).

2 Память DDR4

Процессоры Haswell-EP и Broadwell-EP используют память DDR4. Помимо увеличения пропускной способности DDR4 является более экономичной[4] в сравнении с DDR3. Штатное напряжение питания составляет 1,2 В (для DDR3 — 1,5 и 1,35 В для стандартных и низковольтовых модулей соответственно). Специальных низковольтовых модулей DDR4 пока что не существует.

2.1 Ранги

Поток данных в DDR-SDRAM передаётся блоками по 64 бит (72 бит с учётом ECC). Контроллер памяти выбирает блок такого размера из нескольких чипов DRAM,

[4]См. презентацию Samsung на конференции Memcon 2014 года [9].

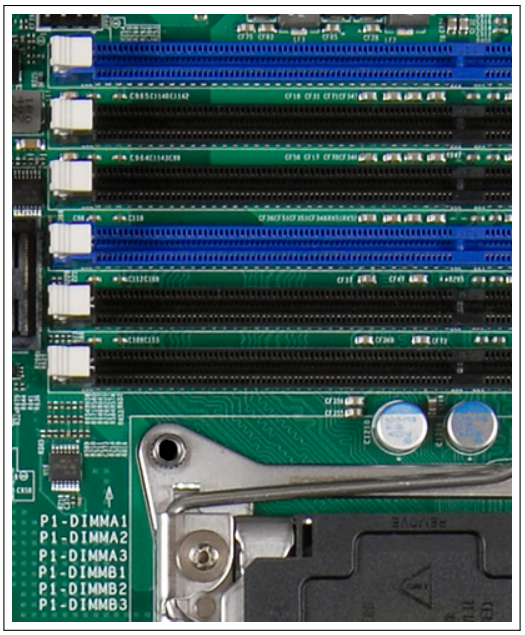

Рис. 2: Маркировка разъёмов DIMM на материнской плате Supermicro X10DRi-LN4+.

каждый из которых отвечает за 4 или 8 бит (**x4** или **x8** в маркировке модулей обозначает *организацию*). Такая группа чипов называется *рангом* (англ. *rank*). Спецификация DDR4 устанавливает ограничение в 8 рангов на канал. Для преодоления этого ограничения были разработаны модули LRDIMM.

Также количество рангов оказывает влияние на производительность — см. раздел 4.2.

2.2 Небуферизованная (UDIMM)

Небуферизованные модули памяти, вне зависимости от наличия поддержки ECC, не поддерживается в системах на базе процессоров Intel Xeon E5-1600/2600 v3/v4.

2.3 Регистровая (RDIMM)

В регистровых модулях памяти (RDIMM) используются дополнительные регистры для буферизации сигналов управления и выставления адреса в целях снижения нагрузки на контроллер памяти. Применение RDIMM позволяет устанавливать большее количество модулей (в случае DDR4 — до 3 модулей на канал) большего объёма[5].

2.4 Память со сниженной нагрузкой (LRDIMM)

В модулях памяти LRDIMM (Load-Reduced DIMM) обеспечивается буферизация не только команд, но и данных. Применение LRDIMM уменьшает электрическую нагрузку

[5]Максимальный объём DDR4 UDIMM составляет 16 ГБ, DDR4 RDIMM — 32 ГБ.

на контроллер памяти в сравнении с RDIMM. Модули LRDIMM используют так называемые *логические ранги* — контроллер памяти определяет количество рангов в 2 раза меньше реального. Например, четырехранговые модули LRDIMM выглядят для контроллера памяти как двухранговые. На практике применение LRDIMM позволяет устанавливать бо́льшее количество памяти (при некотором снижении производительности [6]) или повысить частоту работы памяти.

Модули 3DS TSV LRDIMM (*3 Dimensional Stacked, Through Silicon Vias*) — модули с многослойной компоновкой чипов. По состоянию на начало 2017 года доступны модули объёмом 128 ГБ, в перспективе ожидается появление модулей на 256 ГБ. Особенности архитектуры (данные к буферу передаёт только один чип из всей «стопки» чипов) сочетание в одной системе модулей 3DS TSV LRDIMM и обычных LRDIMM не допускается. Суммарный объём памяти для 2-процессорных материнских плат с 24 разъёмами DIMM может достигать 3 ТБ (24 модуля 128 ГБ 3DS LRDIMM).

3 Базовые правила установки модулей памяти

1. При отсутствии второго процессора в 2-процессорной материнской плате для установки памяти будет доступна только половина разъёмов.
2. Для минимизации потерь производительности желательно обеспечить максимально равномерное размещение модулей по каналам.
3. Небуферизованные модули (UDIMM) **не поддерживаются**.
4. Максимум 8 логических рангов на канал.
5. Смешивать LRDIMM и RDIMM **нельзя**.
6. Смешивать 3DS TSV LRDIMM и обычные LRDIMM или RDIMM **нельзя**.
7. Смешивать память с разной организацией чипов (x4 и x8) **нельзя**.
8. Установка модулей с разным количеством рангов на одном канале возможна.
9. Установка модулей разного объёма на одном канале возможна.
10. Модули с бо́льшим количеством рангов устанавливаются первыми.
11. Установка модулей с различной частотой не рекомендуется.

4 Производительность

Для получения конфигурации с максимально возможной производительностью оперативной памяти следует учитывать множество факторов: выбор процессора по максимально возможной частоте работы памяти (см. таблицы 1, 2, 4, 3), установка модулей с учётом 4-канальности контроллера памяти (балансировка), выбор модулей памяти

(количество рангов и тип модулей), нагрузка на контроллер памяти (количество модулей на канал).

В 2-процессорных системах на производительность доступа к памяти соседнего процессора (между узлами NUMA) оказывает влияние пропускная способность шины QPI.

Влияние на производительность различных технологий оптимизации работы кэша L3 рассматривается в разделе *Optimization of the cache coherence protocol* документа Fujitsu *Fujitsu Server Primergy Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems* [5].

4.1 Балансировка модулей

Для получения сбалансированной конфигурации по всем каналам контроллера памяти и по системе в целом необходимо придерживаться следующих правил:

- Все каналы памяти должны иметь идентичную конфигурацию — общий объём памяти и количество рангов. Количество модулей должно быть кратно 4 на процессор (8 для 2-процессорной системы).
- Все процессоры должны иметь идентичную конфигурацию памяти.

Если количество модулей памяти не позволяет получить полностью сбалансированную конфигурацию, то следует стремиться к как можно более равномерному распределению модулей по каналам.

Примеры:

1. 6 модулей памяти на процессор (12 на два процессора) — см. рис. 3. Первые четыре модуля устанавливаются равномерно по четырём каналам (и для них обеспечивается 4-канальное чередование по каналам), оставшиеся два — в каналы 1 и 2.

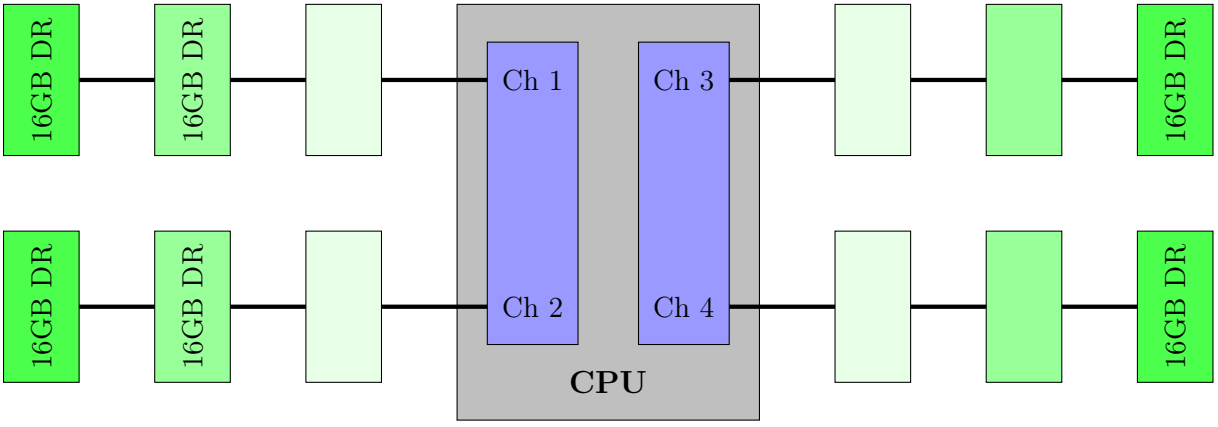

Рис. 3: Максимально сбалансированная установка 6 модулей памяти.

2. 6 модулей памяти на процессор (12 на два процессора) — пример неоптимальной конфигурации, см. рис. 4. Все 6 модулей распределены по двум каналам контроллера. При такой установке для всех модулей обеспечивается лишь 2-канальное чередование, и снижается частота работы памяти из-за установки трёх модулей на канал (в случае процессоров v4 и памяти RDIMM — с 2133 МГц до 1600 МГц, см. раздел 4.3).

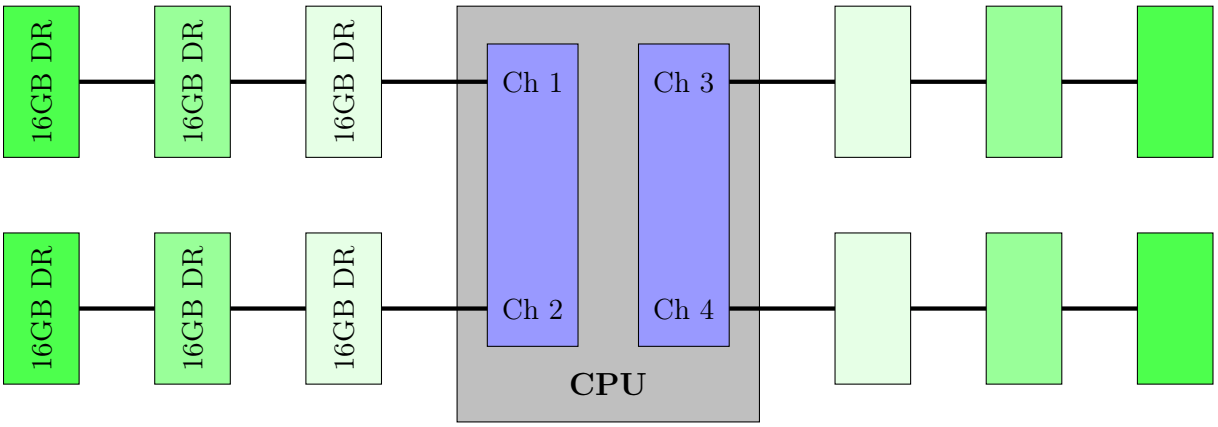

Рис. 4: Несбалансированная установка 6 модулей памяти.

3. Требуется установка 96 ГБ памяти на каждый процессор. Использование шести модулей по 16 ГБ не позволит получить полностью сбалансированную конфигурацию. Существует три варианта решения проблемы:

   - Вместо 6×16 ГБ можно установить 12 модулей по 8 ГБ. Полученная конфигурация будет сбалансирована по каналам (3 модуля по четырём каналам), но память будет работать на меньшей частоте (см. раздел 4.3).
   - Использовать модули разного объёма — 4×16 ГБ и 4×8 ГБ. Восемь модулей можно равномерно распределить по четырём каналам.
   - Вместо 96 ГБ можно установить 128 ГБ меньшим числом модулей. Например, 8×16 ГБ или 4×32 ГБ[6]. Такой вариант обеспечивает наилучшую производительность и оставляет запас для дальнейшего наращивания памяти.

4.2 Влияние количества рангов на производительность

Помимо чередования доступа к памяти по каналам (англ. *channel interleaving*) процессоры E5 могут использовать чередование по рангам (англ. *rank interleaving*). Значительное, порядка 12 %, увеличение пропускной способности памяти достигается при использовании 2-кратного чередования по рангам (установка по одному 2-ранговому

[6]Модули 32 ГБ существуют как в виде обычных RDIMM, так и LRDIMM. Стоит учитывать небольшое отставание LRDIMM в производительности (см. раздел 4.4) за счёт большей гибкости.

модулю на канал или по два 1-ранговых модуля на канал). Переход к 4-кратному чередованию может обеспечить ещё 2-3 %[7].

4.3 Количество модулей на канал

Зависимость максимальной частоты работы памяти от количества установленных модулей представлена в таблицах 5 и 6. Обратите внимание — применение LRDIMM позволяет сохранить частоту в 2400 МГц (2133 МГц для процессоров предыдущего поколения) при установке двух модулей на канал. При этом стоит учитывать меньшую производительность LRDIMM из-за полной буферизации при работе на одинаковой частоте.

В таблицах приведены частоты в соответствии со спецификациями Intel. Некоторые производители серверов, такие как Hewlett Packard [7] и Fujitsu [5] могут использовать более высокие частоты работы памяти (например 2400 МГц вместо 2133 МГц при двух RDIMM на канал).

4.4 Производительность LRDIMM

В модулях памяти со сниженной нагрузкой (LRDIMM) обеспечивается буферизация как команд, так и данных. Побочным эффектом полной буферизации является незначительное увеличение задержки. Поэтому применение LRDIMM может как поднять производительность системы за счёт более высокой частоты работы памяти (например, при установке трёх модулей на канал), так и несколько снизить её в сравнении RDIMM при отсутствии преимущества по частоте работы. Исследования, проведённые компанией Microway показали снижение пропускной способности памяти на 1–7 % в зависимости от режима тестирования и количества модулей на канал [6].

4.5 Cluster on Die

Процессоры E5-2600 с топологиями MCC и HCC (Medium Core Count и High Core Count) — E5-2600 v3 с 10 и более ядрами, E5-2600 v4 с 12 и более ядрами — вместо одного 4-канального контроллера оперативной памяти оснащены двумя 2-канальными контроллерами. Половина ядер использует один контроллер, половина — другой. Доступ к памяти соседнего контроллера осуществляется через внутреннюю шину процессора. Накладные расходы при этом меньше, чем при доступе к памяти другого процессора через QPI, но они всё равно остаются. Операционная система не учитывает эту особенность архитектуры — она видит процессор как один узел NUMA.

[7] См. раздел *Influence of the DIMM types* в документе Fujitsu *Fujitsu Server Primergy Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems* [5].

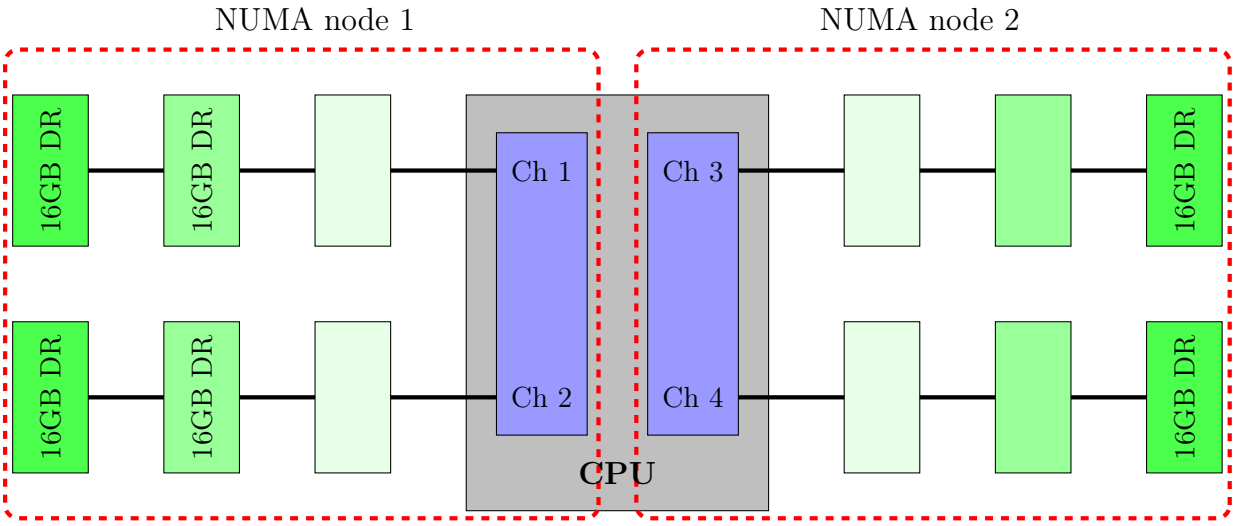

Рис. 5: Cluster on Die в процессорах Intel Xeon E5-2600 v3/v4.

Технология Cluster on Die (CoD, «кластер на кристалле») позволяет разделить процессор на два узла NUMA с учётом принадлежности контроллеров памяти к ядрам. Отключение опции Cluster on Die в настройках BIOS[8] может потребоваться по нескольким причинам:

- Некоторые программные продукты лицензируются по числу процессорных разъёмов, которое раньше всегда соответствовало числу узлов NUMA. При включении CoD происходит удвоение узлов NUMA, что приводит к удвоению стоимости лицензий. Также гипервизор должен учитывать особенности CoD при размещении виртуальных машин с большим числом количеством ядер (больше, чем в одном узле NUMA) — размещать такие ВМ по возможности следует в пределах одного физического процессора. Эти проблемы были устранены в VMware ESXi 5.5 U3b и 6.0 U1 [8].

- Включение CoD в большинстве случаев приводит к увеличению производительности: увеличивается пропускная способность и уменьшается задержка при доступе к памяти локального контроллера. Но при этом на один узел CoD-NUMA будет приходиться в два раза меньший объём кэша L3, и это снизит частоту попадания в кэш [7].

- Включение CoD на системе с неоптимально сбалансированной установкой модулей памяти может обеспечить более равномерную производительность ценой снижения пиковой производительности. В примере на рис. 5 без CoD обеспечивается 4-канальное чередование для 64 ГБ памяти (первые четыре модуля) и 2-канальное для оставшихся 32 ГБ. При включении CoD операционная система будет выделять память для процессов, по мере возможности, на участке, относящемся к узлу NUMA, на котором работают эти процессоры — в данном случае доступ ко всей

[8] В платах Supermicro этот параметр называется *COD Enable*.

памяти будет 2-канальным, но исчезнет задержка, связанная с доступом к памяти «чужого» контроллера[9].

5 Режимы повышения отказоустойчивости

5.1 Mirroring (зеркалирование)

Для включения режима зеркалирования требуется идентичная конфигурация памяти на всех четырёх каналах. При этом модули памяти в двух парах каналов дублируют содержимое друг друга. Такой режим можно рассматривать как аналог дискового массива RAID-1: в сравнении с обычным режимом объём и производительность на запись уменьшаются в 2 раза, производительность на чтение остаётся прежней. В тестах, проведённых компанией Fusjitsu, было получено снижение пропускной способности в тесте STREAM на 23–32 % [5].

5.2 Sparing (резервирование)

При включении резервирования[10] фактически выполняется резервирование на уровне рангов, а не модуля памяти. При обнаружении большого количества корректируемых ошибок в модуле памяти содержимое соответствующего ранга в фоновом режиме копируется в резервный модуль памяти, затем ранг сбойного модуля отключается. Включение резервирования приводит к небольшому снижению пропускной способности памяти — около 7–13 % согласно данным Fujitsu [5].

[9] Ещё один побочный эффект — узлы CoD-NUMA получат разный объём оперативной памяти.
[10] В платах Supermicro эта опция называется *Memory Rank Sparing*.

A Таблицы

Таблица 1: Процессоры E5-2600 v4

Процессор	Кол-во ядер	QPI, ГТ/с	TDP, Вт	Базовая частота, ГГц	Turbo Boost, ГГц	Макс. частота памяти, МГц
E5-2623 v4	4	8,0	85	2,6	3,2	2133
E5-2637 v4	4	9,6	135	3,5	3,7	2400
E5-2603 v4	6	6,4	85	1,7	—	1866
E5-2643 v4	6	9,6	135	3,4	3,7	2400
E5-2609 v4	8	6,4	85	1,7	—	1866
E5-2620 v4	8	8,0	85	2,1	3,0	2133
E5-2667 v4	8	9,6	135	3,2	3,6	2400
E5-2630L v4	10	8,0	55	1,8	2,9	2133
E5-2630 v4	10	8,0	85	2,2	3,1	2133
E5-2640 v4	10	8,0	90	2,4	3,4	2133
E5-2650 v4	12	9,6	105	2,2	2,9	2400
E5-2650L v4	14	9,6	65	1,7	2,5	2400
E5-2660 v4	14	9,6	105	2,0	3,2	2400
E5-2680 v4	14	9,6	120	2,4	3,3	2400
E5-2683 v4	16	9,6	120	2,1	3,0	2400
E5-2690 v4	16	9,6	135	2,6	3,5	2400
E5-2697A v4	16	9,6	145	2,6	3,6	2400
E5-2695 v4	18	9,6	120	2,1	3,3	2400
E5-2697 v4	18	9,6	145	2,3	3,6	2400
E5-2698 v4	20	9,6	135	2,2	3,6	2400
E5-2699 v4	22	9,6	120	2,2	3,6	2400

Таблица 2: Процессоры E5-1600 v4

Процессор	Кол-во ядер	TDP, Вт	Базовая частота, ГГц	Turbo Boost, ГГц	Макс. частота памяти, МГц
E5-1620 v4	4	140	3,5	3,8	2400
E5-1630 v4	4	140	3,7	4,0	2400
E5-1650 v4	6	140	3,6	4,0	2400
E5-1660 v4	8	140	3,2	3,8	2400
E5-1680 v4	8	140	3,4	4,0	2400

Таблица 3: Процессоры E5-1600 v3

Процессор	Кол-во ядер	TDP, Вт	Базовая частота, ГГц	Turbo Boost, ГГц	Макс. частота памяти, МГц
E5-1603 v3	4	140	2,8	—	1866
E5-1607 v3	4	140	3,1	—	1866
E5-1620 v3	4	140	3,5	3,6	2133
E5-1630 v3	4	140	3,7	3,8	2133
E5-1650 v3	6	140	3,5	3,8	2133
E5-1660 v3	8	140	3,0	3,5	2133
E5-1680 v3	8	140	3,2	3,8	2133

Таблица 4: Процессоры E5-2600 v3

Процессор	Кол-во ядер	QPI, ГТ/с	TDP, Вт	Базовая частота, ГГц	Turbo Boost, ГГц	Макс. частота памяти, МГц
E5-2623 v3	4	8,0	105	3,0	3,5	1866
E5-2637 v3	4	9,6	135	3,5	3,7	2133
E5-2603 v3	6	6,4	85	1,6	—	1600
E5-2608L v3	6	6,4	52	2,0	—	1866
E5-2609 v3	6	6,4	85	1,9	—	1600
E5-2620 v3	6	8,0	85	2,4	3,2	1866
E5-2643 v3	6	9,6	135	3,4	3,7	2133
E5-2618L v3	8	8,0	75	2,3	3,4	1866
E5-2630 v3	8	8,0	85	2,4	3,2	1866
E5-2630L v3	8	8,0	55	1,8	2,9	1866
E5-2640 v3	8	8,0	90	2,6	3,4	1866
E5-2667 v3	8	9,6	135	3,2	3,6	2133
E5-2628L v3	10	8,0	75	2,0	2,5	1866
E5-2650 v3	10	9,6	105	2,3	3,0	2133
E5-2660 v3	10	9,6	105	2,6	3,3	2133
E5-2687W v3	10	9,6	160	3,1	3,5	2133
E5-2648L v3	12	9,6	75	1,8	2,5	2133
E5-2650L v3	12	9,6	65	1,8	2,5	2133
E5-2658 v3	12	9,6	105	2,2	2,9	2133
E5-2658A v3	12	9,6	105	2,2	2,9	2133
E5-2670 v3	12	9,6	120	2,3	3,1	2133
E5-2680 v3	12	9,6	120	2,5	3,3	2133
E5-2685 v3	12	9,6	120	2,6	3,3	2133
E5-2690 v3	12	9,6	135	2,6	3,5	2133
E5-2683 v3	14	9,6	120	2,0	3,0	2133
E5-2695 v3	14	9,6	120	2,3	3,3	2133
E5-2697 v3	14	9,6	145	2,6	3,6	2133
E5-2698 v3	16	9,6	135	2,3	3,6	2133
E5-2699 v3	18	9,6	145	2,3	3,6	2133

Таблица 5: Максимальная частота работы памяти в зависимости от количества модулей на канал для Xeon E5-1600/2600 v4

Тип процессора	RDIMM			LRDIMM		
Кол-во модулей на канал	1	2	3	1	2	3
DDR4-2400	2400	2133	1600	2400	2400	1866
DDR4-2133	2133	2133	1600	2133	2133	1866
DDR4-1866	1866	1866	1600	1866	1866	1866

Таблица 6: Максимальная частота работы памяти в зависимости от количества модулей на канал для Xeon E5-1600/2600 v3

Тип процессора	RDIMM			LRDIMM		
Кол-во модулей на канал	1	2	3	1	2	3
DDR4-2133	2133	1866	1600	2133	2133	1600
DDR4-1866	1866	1866	1600	1866	1866	1600
DDR4-1600	1600	1600	1600	1600	1600	1600

B Изменения в документе

- **24.03.2017**. Скорректированы данные по поддержке процессорами технологии Cluster on Die (см. раздел 4.5).

Список литературы

[1] *Memory Configuration Guide: X10 Series DP Motherboards – (Socket R3)*, (Supermicro, 2014): goo.gl/qo59vT.

[2] Dan Colglazier, Joseph Jakubowski, Tristian Brown, *Maximizing System Performance with a Balanced Memory Configuration*, (Lenovo, 2016): goo.gl/qj0rlN.

[3] David Mulnix, *Intel Xeon Processor E5-2600 v4 Product Family Technical Overview*, (Intel, 2016): goo.gl/rCO5vJ.

[4] *FUJITSU Server Primergy Memory Performance of Xeon E5-2600 v3 (Haswell-EP) based Systems*, (Fujitsu, 2015): goo.gl/T6iAWw.

[5] *Fujitsu Server Primergy Memory Performance of Xeon E5-2600 v4 (Broadwell-EP) based Systems*, (Fujitsu, 2016): goo.gl/eN1jWk.

[6] Marc Rocque, *DDR4 RDIMM and LRDIMM Performance Comparison*, (Microway, 2016): goo.gl/CdYQDV.

[7] *Overview of DDR4 memory in HPE ProLiant Gen9 Servers with Intel Xeon E5-2600 v3*, (Hewlett Packard Enterprise, 2015): HPE website.

[8] *Intel Cluster-on-Die (COD) Technology, and VMware vSphere 5.5 U3b and 6.x (2142499)*, (VMware, 2016): goo.gl/ODP9nn.

[9] *Understanding DDR4 and Today's DRAM Frontier*, (Samsung, 2014): goo.gl/Aljvwr.

[10] Frank Denneman, *NUMA Deep Dive Part 2: System Architecture*, (2016): goo.gl/XPTkxc.

[11] Frank Denneman, *NUMA Deep Dive Part 3: Cache Coherency*, (2016): goo.gl/zEgsc1.