ОЦЕНКА ВОЗМОЖНОЙ ОПТИМИЗАЦИИ СЕТИ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ КЛАСТЕРНЫХ ВЫЧИСЛЕНИЙ (GIGABIT ETHERNET)

Ахмед Набиль Мухаммед Мудхш
Аспирант, Санкт-Петербургский государственный электротехнический университет СПбГЭТУ «ЛЭТИ»
Aboroan1987@yahoo.com

Аннотация. В работе представлены общие сведения о вычислительных кластерах, а так же факторы влияющие на производительность вычислительных кластеров. Программные приложения для кластера. Оценка верхней границы для возможности оптимизации работы сетевого стека на используемом оборудовании.

Ключевые слова: оптимизация, высокопроизводительные вычисления, Gigabit Ethernet, сетевой стек, кластер.

THE EVALUATION OF THE POSSIBLE OPTIMIZATION OF THE NETWORK OF HIGH-PERFORMANCE CLUSTER COMPUTING (GIGABIT ETHERNET)

N. Ahmed

Summary. In work it is presented the general information about computing cluster, Factors that influence on productivity of computing cluster. Application program for cluster. An assessment of the top border for a possibility optimization works of a network stack on the used equipment.

Keywords: optimization, high performance computing, Gigabit Ethernet, network stack, cluster.

Оборудование кластера

В рамках данной статьи предполагается оптимизация используемого программного стека при неизменном оборудовании. Так, целевой вычислительный кластер, для которого предлагаются решения, дан и не изменяется в рамках предлагаемых оптимизаций. Это типичная ситуация для вычислительного центра: оборудование закупленного кластера редко обновляется, так как кластер рассчитан на определенную нагрузку. При появлении существенно новых задач, при существенном изменении требований (например, резкое увеличение числа пользователей), как правило, выполняется покупка нового современного кластера. Это оправдано быстрым устареванием оборудования в рассматриваемой области. Покупка нового кластера требует значительных денежных средств. Однако существует вариант оптимизации работы существующего кластера без обновления оборудования. Этот вариант подходит как решение для случая, когда требуется ускорение расчетов, однако, существенного изменения требований к кластеру нет. Такая ситуация может возникнуть в процессе эксплуатации кластера. Тем более что обновление существующего кластера не всегда простое решение: с учетом скорости устаревания оборудования и появления новых технологий и стандартов поиск оборудования для старого кластера может быть связан с трудностями. Подобное обновление может также потребовать существенных денежных средств. Таким образом, выбор подходящих алгоритмов и протоколов, оптимизация работы программного обеспечения (в том числе операционной системы), оптимизация использования оборудования, его подходящая конфигурирование могут повысить эффективность расчетов без необходимости в затратах на оборудование. Такой оптимизации и посвящена данная работа.

С учетом вышесказанного необходимо детально рассмотреть оборудование вычислительного кластера — типичный узел кластера. Это необходимо для оценки перспектив указанных оптимизаций. Будут рассмотрены серверный узел (кластер ПМ-ПУ) и десктопный узел, а также предлагаемый узел с 10 Гбит Ethernet сетевой картой (об этом будет сказано дальше).

Основное внимание будет уделено скорости обмена данными между различными элементами узла кластера и между отдельными узлами.

На рисунке 1 представлена схема серверного узла.

Узел предусматривает установку двух процессоров архитектуры x86_64 Intel Xeon E5410 (4 ядра, тактовая частота ядра — 2.3 ГГц). Взаимодействие процессора с периферийными устройствами осуществляется благодаря «северному» (Intel 5000 Memory Controller Hub) и «южному» (Intel I/O Controller Hub ESB2-E) мостам. «Северный» мост обеспечивает взаимодействие процессора с оперативной памятью и рядом других высокоскоростных устройств. Скорость соединения процессора с «север-

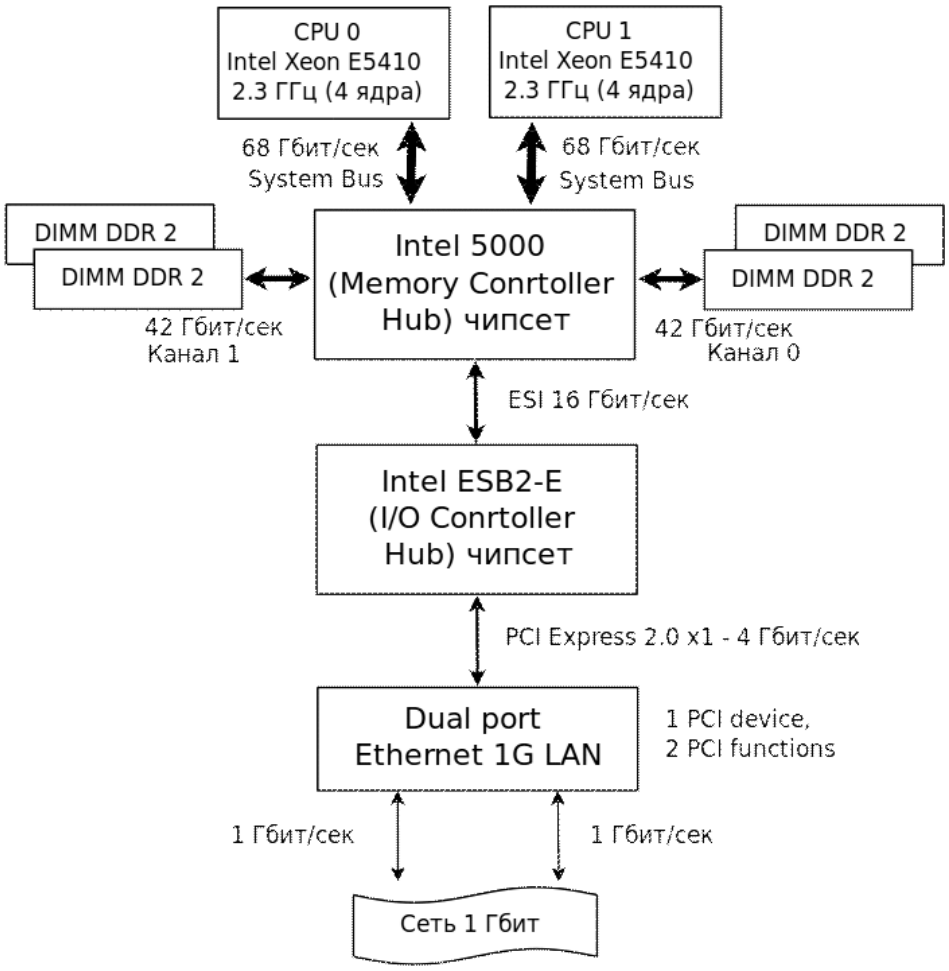

Рис. 1. Схема серверного узла (кластер ПМ-ПУ)

ным» мостом — 68 Гбит/сек. Скорость обмена данными между «северным» мостом и оперативной памятью — 42 Гбит/сек (контроллер памяти поддерживает 2 канала).

Сетевая карта PCI Express подключена к «южному» мосту. «Южный» мост соединен с «северным» мостом по шине ESI, обеспечивающей скорость 16 Гбит/сек. Наконец, сама карта является устройством PCI Express версии стандарта 2.0, она подключена по соединению x1. По сути, это два симплексных канала (lane). Базовая проивзодительность для каждого из направлений — 5 ГТ/сек (гигатранзакций в секунду). В рамках одной транзакции передается один бит, поэтому скорость передачи соответствует скорости 5 Гбит/сек. Однако для борьбы с ошибками используется кодирование 8b/10b, таким образом, скорость передачи полезных данных равна 4 Гбит/сек.

Как видно, скорости «внутренних» соединений достаточно для обеспечения соединения 1 Гбит/сек стандарта Ethernet 1 Gbit. Однако скорость самого сетевого соединения (между узлами кластера) значительно меньше остальных приведенных скоростей. Так, скорость обращения к памяти составляет 42 Гбит/сек, в то время как скорость сетевого соединения всего 1 Гбит/сек.

Усугубляет ситуацию и тот факт, что в системе имеется два моста — «северный» и «южный». Современные процессоры x86 (x86_64) уже не имеют «северного» моста. Он увеличивал задержки при обращении процессора к памяти. Сейчас вся его логика перенесена в процессор.

Для сравнения архитектура десктопных узлов значительно отличается от архитектуры серверных узлов. В рамках десктопных узлов используются мобильные процессоры — Intel N3700 (4 ядра, тактовая частота ядра 1.6 ГГц). Это довольно новые процессоры. Архитектуру десктопных узлов отличает тот факт, что в рамках нее уже не используется «северный» мост. Большая часть упомянутой функциональности интегрирована в сам процессор (SoC). Схема десктопного узла приведена на рисунке 2.

Контроллер памяти поддерживает 2 канала и обеспечивает передачу данных со скоростью 102 Гбит/сек.

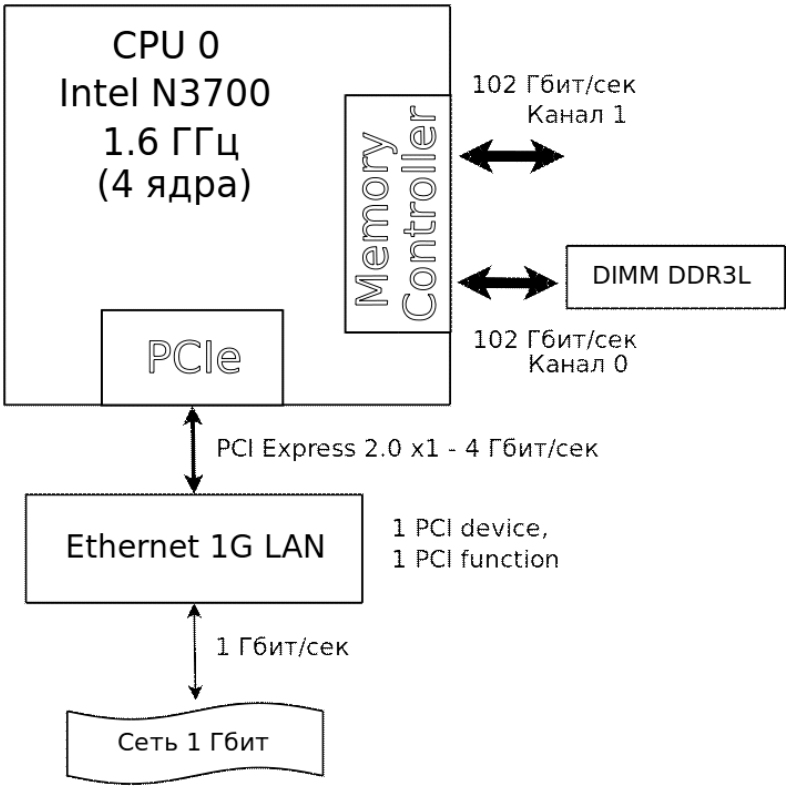

Рис. 2. Схема десктопного узла

Используемая память — форм-фактора DIMM стандарта DDR3L. Сетевая карта — устройство PCI Express версии стандарта 2.0, она подключена по соединению x1. Таким образом, здесь также обеспечивается скорость передачи полезных данных 4 Гбит/сек. Сетевая карта — карта Realtek RTL8111. В рамках Linux за работу с картой отвечает драйве «r8169» (drivers/net/ethernet/realtek/r8169.c).

И здесь снова подобная ситуация: скорости «внутренних» соединений намного больше 1 Гбит/сек стандарта Ethernet 1 Gbit. Так, скорость обращения к памяти на 2 порядка выше скорости соединения по сети. Таким образом и здесь передача данных по сети может стать существенным препятствием для достижения высокой производительности. Именно поэтому оптимизация доступных компонентов (например, программного обеспечения) так важна.

В качестве примера узла с сетью 10 Гбит/сек можно привести узел, схематично изображенный на рисунке 3.

Это NUMA-система, в рамках которой возможна установка 2 процессоров. На схеме, изображенной на рисунке 3 установлен лишь один процессор (семейства Intel Xeon E5 V3). В рамках указанной архитектуры вместо «южного» моста используется Platform Controller Hub (отчасти «похожий» на традиционный «южный» мост). Функциональность «северного» моста интегрирована в сам процессор. Два возможных процессора в системе соединены при помощи шины QPI (скорость до 76 Гбит/сек в рамках данной системы).

Контроллер памяти процессора поддерживает до 4 каналов, скорость — 102 Гбит/сек. Используемый тип памяти — DIMM DDR4. Сетевая карта — устройство PCI Express версии стандарта 2.0, она подключена по соединению x8. Таким образом, при базовой скорости 5 ГТ/сек и кодировании 8b/10b здесь обеспечивается скорость передачи полезных данных 32 Гбит/сек. Сетевая карта — двухпортовый адаптер Intel X540-T2. В рамках Linux за работу с картой отвечает драйвер «ixgbe» (drivers/net/ethernet/intel/ixgbe). Можно обратить внимание на то, что высокоскоростная сетевая карта (10 Гбит) подключена напрямую (мост PCI Express расположен в процессоре), в то время как низкоскоростная карта (1 Гбит) подключена через PCH.

Теперь можно рассмотреть типичные вопросы, возникающие в процессе работы вычислительного центра. Разумеется, перед покупкой нового кластера проводятся оценки для определения нужного числа узлов, оборудования узлов и т.д. Тоже самое относится и к сети — рассчитывается требуемая пропускная способность сети, учитывая задачи кластера.

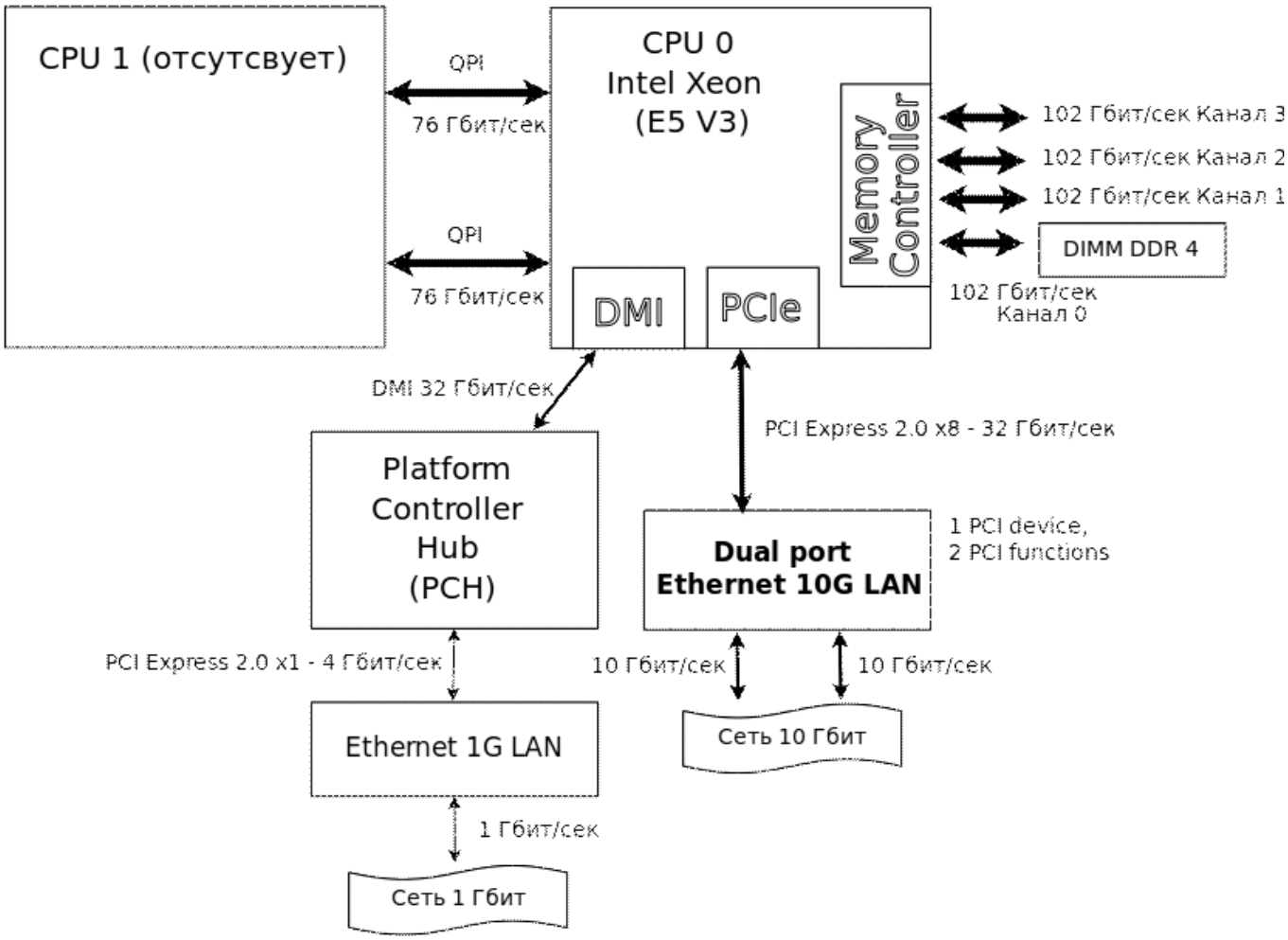

Рисунок 3. Схема узла с 10 Гбит Ethernet

Однако в рамках вычислительного центра подобные расчеты оправдываются далеко не всегда. Требования пользователей меняются довольно быстро, как и потоки информации в рамках вычислительного кластера. Так, для кластера, обеспечивающего работу, к примеру, веб-портала нагрузка и тип задач все время остаются примерно одинаковыми до тех пор, пока портал не приобретет большую популярность или существенно изменит предоставляемые сервисы.

С учетом этих особенностей можно заранее рассчитать и требуемое оборудование, и требуемую производительность сети (исключительные ситуации — отдельная тема для рассмотрения). Также и требования к оборудованию, как правило, ограничиваются рассмотренными вопросами (к слову, оно может быть относительно дешевым — к примеру, векторные расширения процессора для работы с числами с плавающей запятой могут оказаться совершенно невостребованными). Для вычислительного кластера ситуация иная. Множество пользователей запускают расчеты с использованием различных приложений, имеющих различные требования и к оборудованию узлов, и к сети. Конечно, некоторые вычислительные кластеры полностью выделены под одно конкретное приложение. В таких случаях исследователи стараются создать некоторые ad hoc решения лишь для конкретных задач. Но даже в рамках одного приложения может быть реализовано множество алгоритмов, от входных данных которых могут меняться требования к сети (пример — OpenFOAM со множеством солверов).

Как было показано выше (оборудование кластера), производительность работы сети 1 Гбит практически на два порядка отличается от производительности оперативной памяти и производительности самого процессора. Таким образом, очевидна большая разница в работе вычислителя и сетевой подсистемы. Расчет программного стека выполняется на вычислителе (процессоре) и таким образом обработка пакетов выполняется с гораздо большей скоростью, чем их пересылка по сети.

В общем случае для оценки времени обработки пакета и его отправки можно привести следующую формулу:

$$T = C / VC + L1 / VL1 + L2 / VL2 + M1 / VM1 + M2 / VM2 + N / VN + R$$

Здесь

C — число операций, выполняемых процессором (в рамках обработки пакета);

VC — скорость работы процессора;

$L1$ — число обращений процессора к памяти, которые были удовлетворены за счет кэш-памяти 1 уровня;

$VL1$ — скорость работы кэш-памяти 1 уровня;

$L2$ — число обращений процессора к памяти, которые были удовлетворены за счет кэш-памяти 2 уровня;

$VL2$ — скорость работы кэш-памяти 2 уровня;

$M1$ — число обращений к оперативной памяти со стороны процессора;

$VM1$ — скорость такого доступа (могут вноситься задержки «северным» мостом в случае его использования (в рамках старых x86 узлов));

$M2$ — число обращений к оперативной памяти со стороны сетевой карты (по DMA);

$VM2$ — скорость такого доступа (могут вноситься задержки «северным» и «южным» мостами в случае их использования (в рамках старых x86 узлов));

N — число операций по пересылке данных сетевой картой;

VN — скорость работы сетевой карты;

R — время на выполнение прочих операций (например, обработка прерывания не от сетевой карты, ожидание доступа к шине, отсутствие записи в TLB для требуемой страницы и т.п.).

В случае серверных узлов «разрыв» между скоростью работы сетевой карты и скоростью работы остальных подсистем, вовлеченных в обработку пакета довольно велик, к примеру, соотношение между скоростями доступа к памяти со стороны процессора и отправки данных по сети равно:

$$\frac{V_{M1}}{V_N} = 42$$

Скорость доступа к памяти со стороны сетевой карты (по DMA) ограничена участком с самой низкой скоростью и аналогичное соотношение составляет (однако для карты 1 Гбит обеспечиваемых 4 Гбит более чем достаточно):

$$\frac{V_{M2}}{V_N} = 4$$

Для десктопного узла первое соотношение будет равно:

$$\frac{V_{M1}}{V_N} = 102$$

И второе соотношение будет также равно:

$$\frac{V_{M2}}{V_N} = 4$$

Но в данном случае второе соотношение, по сути не важно — требуется лишь выполнение условия:

$$VM2 > VN$$

Ведь в противном случае использование более скоростного сетевого адаптера не имеет смысла, так как скорость пересылки данных между узлами будет ограничена скоростью пересылки данных в рамках одного узла.

Разумеется, разница в скоростях может быть не заметна в случае, к примеру, большого числа операций, выполняемых процессором — в таком пересылка данных по низкоскоростной сети будет составлять лишь небольшую долю в общем времени обработки и отправки пакета (основную часть времени будет составлять обработка пакета на процессоре).

Обработка пакета в рамках теста «ping-pong», к примеру, в рамках десктопного узла в случае использования стандартных средств, составляет всего 7% от общего времени выполнения теста. Остальные 93% — пересылка данных по низкоскоростной сети, получение данных сетевой картой из памяти и работа самой сетевой карты. Таким образом, оптимизация работы сетевого стека приведет лишь к ускорению работы в рамках 7% от общего времени вычислений. Даже если бы удалось свести время работы программного стека к 0, это бы лишь убрало 7%, оставив 93% времени работы. Обозначив время работы существующего используемого программного стека за TS, а время пересылки данных за TH, придем к соотношению, дающему верхнюю границу для оптимизации работы используемого программного стека:

$$L = \frac{T_S}{T_S + T_H}$$

По сути L — это процент увеличения производительности, который можно достигнуть за счет оптимизации работы программного стека. Поскольку в рамках данной работы рассматривается уже имеющееся оборудование (TH неизменен для определенной задачи пересылки данных), максимальное увеличение производительности ограничено этим значением. И для имеющегося оборудования это всего лишь 7%.

В случае использования более нового оборудования с сетевой картой 10 Гбит, L может быть равен 42%. Верхняя граница для случая 1 Гбит (рассматриваемого в данной работе) и для случая 10 Гбит (предлагаемого оборудования) изображены на рисунке 4.

Как видно, в случае использования сети 10 Гбит вместо 1 Гбит можно добиться в 6 раз большего ускорения (в процентах) при использовании тех же самых методов оптимизации работы сети.

Поэтому несмотря на ограничение в 7% для сети 1 Гбит разработка методов оптимизации работы сети может дать заметный прирост производительности отдельно взятого потока данных в случае использования сети 10 Гбит при том же самом остальном оборудовании, то есть в случае выравнивания скорости работы сети и процессора.

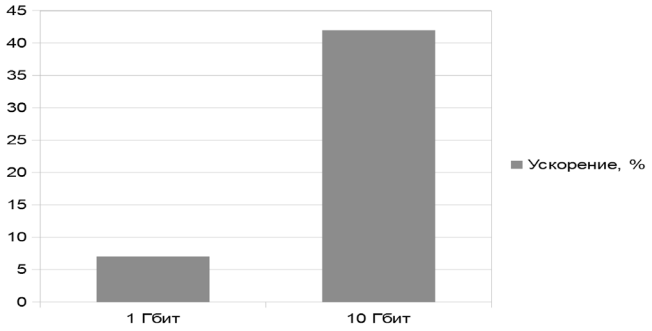

Рис. 4. Верхняя граница для оптимизации.

ЛИТЕРАТУРА

1. Top 500 supercomputers sites (электронных ресурс). — URL: http://www.top500.org/.
2. Воеводин В.В., Жуматий С. А. Вычислительное дело и кластерные системы. — М.: Изд- во МГУ, 2007.
3. Бройдо В Л., Ильина О. П. Архитектура ЭВМ и систем: Учебник для вузов. — 2-е изд. — СПб.: Изд-во Питер, 2009.
4. MPI Performance Measurements, http://www.llnl.gov/computing/mpi/mpi_benchmarks.html.


Санкт-Петербургский государственный электротехнический университет СПбГЭТУ «ЛЭТИ»