

УДК 004.7.057.4

Формат пакета ARTCP. Особенности формирования и обработки заголовков ARTCP в сетевой подсистеме ОС Linux 2.6.

Сивов А.А.

Ярославский государственный университет им. П.Г. Демидова

e-mail: mm05@mail.ru

получена 4 апреля 2011

Ключевые слова: протокол ARTCP, Linux, транспортные протоколы, измерение времени

Описывается формат заголовка ARTCP. Рассматриваются практические аспекты формирования заголовка и вычисления значения его полей. Обсуждаются вопросы точного вычисления времени прибытия сегментов ARTCP. Также в статье описываются интерфейсы ядра Linux для вычисления времени и концепция “источников времени” и ее реализация.

ARTCP (Adaptive Rate TCP) – коммуникационный протокол транспортного уровня, гарантирующий доставку данных. Этот протокол разработан на основе TCP, но в отличие от него использует темпоральные характеристики потока передаваемых данных для управления этим потоком. Основной особенностью ARTCP является логическое разделение механизма коррекции ошибок передачи и управления потоком. Идеи, лежащие в основе ARTCP, и описание протокола рассматриваются в [1] и [2].

Данная статья имеет целью описание формата сообщений ARTCP. В ней рассматривается формирование полей заголовка ARTCP пакета, обсуждаются практические вопросы вычисления их значений. В статье [3] указано, что ОС Linux рассматривается в качестве системы, для которой будет производиться реализация ARTCP. Авторы данной статьи обсуждают ряд изменений, которые необходимо внести в сетевую подсистему ОС Linux для поддержки протокола ARTCP.

1. Формат пакета ARTCP

ARTCP разрабатывался с учетом совместимости с TCP [4]. Например, одной из задач ставилось автоматическое переключение ARTCP соединения в режим TCP (т.е. чтобы оно фактически становилось TCP соединением), если один из концов соединения не поддерживает ARTCP, но поддерживает TCP. (Подробнее данное

Таблица 1. Заголовок TCP

Бит	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1															
0	Порт отправителя																Порт получателя																														
32	Номер последовательности																																														
64	Номер подтверждения																																														
96	Смещение дан- ных				Резерв				Флаги								Окно																														
Контрольная сумма																	Указатель важных данных																														
128																																															
160	Опции (необязательное поле)																																														

переключение при установлении ARTCP соединения рассматривается в статье [3].) Такой подход, безусловно, повлиял на выбор формата пакета ARTCP данных.

Заголовок сегмента ARTCP полностью совместим с заголовком TCP (см. таблицу 1). Помимо полей заголовка TCP он содержит два собственных поля: TI (Time Interval) и PS (Packet Sequence), которые оформлены в заголовке как опции заголовка сегмента TCP. Каждое из этих полей имеет размер 4 байта, но будучи представленным в виде опции TCP, занимает в заголовке по 8 байт каждое (1 байт — номер опции, 1 байт — размер опции, 4 байта — значение опции и 2 байта — выравнивание до следующего 32-битного слова). В настоящее время для поля PS используется номер опции 253, а для поля TI — 254. Эти номера выбраны согласно RFC 4727 [5] для использования в экспериментах и должны быть в дальнейшем заменены в соответствии с RFC 2780 [6]. Использование номеров опций TCP 253 и 254 регламентируется в RFC 3692 [7].

Таблица 2. Поле PS в заголовке ARTCP

Байт	0	1	2	3	4	5	6	7
Значение	1	1	253	6	Значение поля PS			

Таблица 3. Поле TI в заголовке ARTCP

Байт	0	1	2	3	4	5	6	7
Значение	1	1	254	6	Значение поля TI			

Таблицы 2 и 3 иллюстрируют представление полей PS и TI в заголовке сегмента ARTCP в виде опций TCP. ARTCP не регулирует порядок представления

данных полей, но в разрабатываемой нами реализации ARTCP мы придерживаемся следующих правил при формировании заголовка: поля PS и TI располагаются после опций, соответствующих TCP (за исключением опции с номером 0, означающей конец списка опций); если в заголовке присутствуют оба поля PS и TI, то поле PS располагается первым. При обработке принятого сегмента ARTCP мы не делаем предположений о положении полей PS и TI и опций TCP, допуская их произвольную очередность.

Поле TI используется протоколом ARTCP для вычисления значения скважности потока. Изначально описание протокола ARTCP [1] предполагало, что получатель записывает в поле TI значение скорости потока, вычисленное им по формуле $R = S / t$, где S – размер принятого кадра канального уровня, а t – временной промежуток, измеренный между последовательными моментами прибытия сегментов. Данный подход имеет ряд недостатков. Он требует оперирования с размером кадра канального уровня, что противоречит идее модели OSI рассматривать нижележащий протокол как абстракцию, предоставляющую определенные сервисы. Более того, использование такой величины, как размер кадра канального уровня, для вычисления темпоральных характеристик потока в гетерогенных сетях представляется некорректным, т.к. кадр, содержащий один и тот же сегмент ARTCP, может иметь разный размер на разных участках сети в связи с различными протоколами канального уровня, используемыми на этих участках. В связи с этим в статье [3] предлагается использовать вместо размера кадра канального уровня размер полезных данных принятого сегмента ARTCP.

Еще одним недостатком является вычисление скорости потока на стороне получателя, а не на стороне отправителя. Это требует передачи по сети значений с плавающей точкой, что может быть нежелательно, и затрудняет интерпретацию отправителем полученных данных в случае, если временной промежуток необходимо скорректировать.

В статье [3] предлагается передавать в поле TI временной промежуток, измеренный между последовательными моментами прибытия сегментов ARTCP. При таком подходе отправитель, получив эти данные, способен скорректировать их и вычислить скорость потока самостоятельно. Отправитель знает размер переданного им сегмента ARTCP и имеет больше информации о переданных сегментах, чем получатель. В частности, он знает время отправки каждого из сегментов, был ли сегмент отправлен позже, чем это позволяют характеристики сети (например, из-за отсутствия данных у отправителя), и поэтому, имея предоставленную получателем информацию в состоянии, лучше оценить скорость потока.

В качестве размерности временного промежутка, передаваемого в поле TI, необходимо выбрать фактическую единицу времени, не зависящую от аппаратного обеспечения, используемого отправителем или получателем (т.е. эта единица не может вычисляться в тактах процессора или других подобных величинах). Статья [3] предлагает использовать для данной цели микросекунды. Измерение времени с данной размерностью возможно на большинстве используемых в настоящее время аппаратных архитектур и существующих процессоров, а также обладает достаточной точностью, необходимой для поставленной задачи. (Практические аспекты точного вычисления времени рассматриваются в данной статье далее.)

Приведем пример, подтверждающий, что применение данной размерности вычисления времени обладает достаточной точностью. Предположим, что два компьютера соединены между собой каналом с пропускной способностью 1 Гбит/сек и один из компьютеров отправляет другому пакеты размером 1 Кбайт. Очевидно, что за одну секунду может быть передано не более 125000 пакетов, т.е. 1 пакет за 8 мкс.

Поле PS служит для передачи порядкового номера сегмента в направлении от отправителя к получателю. В описании ARTCP, представленном в [1], сказано, что получатель использует значение этого поля, чтобы определить, что принятый им пакет является следующим в потоке по отношению к предыдущему принятому им пакету. Т.е. значение поля PS в протоколе ARTCP служит тому, чтобы получатель мог отличить, передан ли пакет, полученный им, отправителем первый раз или повторно. Действительно, данных, содержащихся в заголовках сегментов TCP, достаточно, чтобы различать сегменты, переданные в неправильном порядке, используя для этого номер последовательности, но с точки зрения TCP сегмент, переданный отправителем первый раз, ничем не отличается от того же сегмента, переданного им повторно (в случае, если сегмент был “потерян” в сети), т.к. он имеет тот же номер последовательности и не имеет никакого индикатора, говорящего, что происходит повторная передача.

Каждый сегмент ARTCP, содержащий полезные данные, а также сегмент, содержащий флаг SYN или флаг FIN, содержит в своем заголовке поле PS, в котором передается номер сегмента. При этом данный номер на единицу больше номера предыдущего сегмента, переданного отправителем (за исключением повторной передачи сегментов при установке соединения). Т.е. если сегмент со значением N в поле PS был передан отправителем, но не был получен получателем (или отправитель не получил уведомление), то тот же сегмент передается повторно, но уже со значением N+1 в поле PS (очевидно, что N+1 берется по модулю 2^{32}). Каждый сегмент ARTCP, содержащий флаг ACK, должен содержать в заголовке поле TI. В случае, если уведомление происходит в ответ на сегмент, содержащий в поле PS значение, отличающееся от значения, на единицу большего по модулю 2^{32} значения поля PS предыдущего принятого сегмента, в поле TI передается 0, в противном случае в поле TI передается вычисленное значение интервала времени. Значения полей PS и TI записываются в сетевом порядке.

При установке соединения ARTCP в полях TI и PS соответствующих пакетов передается значение 0. Подробнее установка соединения рассмотрена в статье [3].

2. Реализация чтения и записи заголовков сегментов ARTCP в сетевой подсистеме ОС Linux

Протокол ARTCP заимствует многие механизмы из протокола TCP, к тому же реализация ARTCP должна позволять переключаться на использование TCP в случае, если один из участников соединения не поддерживает ARTCP, поэтому нами было решено при реализации протокола ARTCP для ОС Linux использовать существующую сетевую подсистему данной операционной системы и, в частности, реализацию стека IPv4/TCP.

Как было сказано ранее, заголовок сегмента ARTCP отличается от заголовка сегмента TCP исключительно наличием полей PS и TI, представленных в виде опций TCP, поэтому для реализации записи и чтения заголовка ARTCP необходимо модифицировать код, реализующий запись и чтение опций заголовка TCP. Данный код расположен в файлах `/net/ipv4/tcp_input.c` и `/net/ipv4/tcp_output.c` дерева исходных кодов ядра Linux.

Для формирования опций, которые необходимо отправить в заголовке, в Linux используется структура `struct tcp_out_options`, содержащая поля со значениями различных опций TCP, поддерживаемых сетевым стеком Linux, и битовое поле `options`, в котором устанавливаются флаги, означающие, какие опции нужно записать в заголовок данного конкретного сегмента TCP. Для реализации записи заголовка ARTCP в данную структуру были добавлены поля, вмещающие значения TI и PS, и были введены битовые флаги, указывающие на наличие этих полей и использующиеся в битовом поле `options`.

Помимо этого были внесены изменения в функции, формирующие экземпляр структуры `struct tcp_out_options` и записывающие данные в заголовок сегмента на основе переданного в функцию экземпляра данной структуры. Для формирования экземпляра структуры в сетевой подсистеме Linux используются функции `tcp_syn_options`, `tcp_synack_options` и `tcp_established_options`, формирующие опции TCP для SYN пакета, SYN-ACK пакета и всех остальных пакетов, соответственно. Реализация данных функций для ARTCP добавляет запись в структуру информации о полях PS и TI, если сокет находится в режиме поддержки ARTCP.

Для записи опций TCP в буфер, содержащий заголовок, используется функция `tcp_options_write`. В данную функцию также были внесены изменения, чтобы она записывала поля PS и TI в буфер заголовка сегмента (в виде, указанном в таблицах 2 и 3), если это указано в переданном ей экземпляре структуры `struct tcp_out_options`. Эти изменения позволяют формировать сегмент ARTCP в сетевой подсистеме Linux.

Для разбора поля опций заголовка принятого сегмента TCP в ОС Linux вызывается функция `tcp_parse_options`, которая анализирует принятые данные и формирует структуру `struct tcp_options_received`, записывая в неё полученные значения. Для поддержки ARTCP в эту структуру были добавлены поля `ps` и `ti`, вмещающие значения полей PS и TI заголовка рассматриваемого сегмента ARTCP, и битовое поле `artcp_options`, определяющее, какие из этих двух полей действительно присутствовали в заголовке (ни одного, поле PS, поле TI или оба). Также был модифицирован код функции `tcp_parse_options`, чтобы он мог обрабатывать поля ARTCP и формировать измененную структуру.

Помимо разбора полей заголовка ARTCP для обработки пришедшего пакета необходима обработка данных полей в зависимости от состояния соединения и наличия/отсутствия полезных данных в полученном сегменте. Полученный SYN пакет обрабатывается в случае стека IPv4/TCP в функции `tcp_v4_conn_request`. Модифицированный код данной функции выполняет проверку настроек ОС, установленных администратором, считывая значение переменной `sysctl_tcp_enable_artcp`, которая указывает, включен ли ARTCP глобально в системе, и, если ARTCP включен, проверяет наличие поля PS и корректность его значения (а также отсут-

ствие поля TI). Если проверки выполнены, то сокет переводится в режим работы ARTCP и проводится инициализация необходимых для ARTCP соединения ресурсов. В противном случае сокет переводится в режим работы TCP.

Если сокет уже отправил SYN пакет (и находится в состоянии SYN-SENT), то принимаемые им пакеты обрабатываются функцией `tcp_rcv_synsent_state_process`. Модифицированный код данной функции в случае получения SYN-ACK пакета для сокета, находящегося в режиме ARTCP, проверяет наличие в заголовке полей PS и TI и корректность их значения. Если проверки не выполняются, то сокет переводится в режим TCP. В противном случае проводится инициализация необходимых для ARTCP соединения ресурсов.

Если ARTCP соединение установлено, то принимаемые пакеты обрабатываются функцией `tcp_rcv_established`. Для сокета, находящегося в режиме ARTCP, модифицированный код данной функции проводит обработку сегментов ARTCP. Если в заголовке принятого пакета отсутствуют необходимые поля PS (для сегмента с полезными данными) или TI (для сегмента с флагом ACK) или в заголовке сегмента, не содержащего полезных данных, имеется поле PS, то сокет переводится в режим TCP. Если принят сегмент с флагом ACK, то значение поля TI передается в функцию управления потоком ARTCP. При обработке сегмента, содержащего полезные данные, проверяется значение поля PS. Если, согласно значению этого поля, данный сегмент является следующим (после предыдущего принятого сегмента) сегментом, отправленным другим концом соединения, то необходимо вычислить разность между временем прибытия этих двух сегментов, чтобы передать её в поле TI ответного ACK пакета. Иначе в поле TI следует передать 0.

3. Практические аспекты высокоточного вычисления времени

Как сказано ранее, поле TI требует вычисления интервалов времени в микросекундах. Вычисление интервалов времени с необходимой точностью может оказаться нетривиальной задачей. Ядро Linux гарантирует наличие т.н. “системного таймера”, доступного через интерфейс `jiffies` и обновляемого с частотой **HZ**, задаваемой при компиляции ядра [8]. Разумный диапазон значений **HZ** составляет от 100 до 1000 Гц, что соответствует точности вычисления интервалов 1-10 мс.

Наличие более точного способа вычисления интервалов времени зависит от аппаратного обеспечения, на котором запущена ОС Linux. Рассмотрим в качестве примера архитектуру x86. Все IBM-совместимые компьютеры имеют программируемый таймер интервалов (Programmable Interval Timer, PIT), известный как чип Intel 8253. Этот чип или его аналог (в современных материнских платах данную функциональность предоставляет “южный мост”) имеет три независимых канала счета с 16-разрядными счетчиками. Обычно 0-й канал используется для генерации прерываний часов, 1-й — для работы с DRAM памятью, а 2-й — для генерации тонов динамика ПК (PC speaker). Использование данного таймера позволяет добиться точности 1 мс.

Другим источником времени являются часы реального времени (RTC). Функции

ональность данной микросхемы в современных материнских платах также предоставляет “южный мост”. Как и PIT, RTC предоставляет точность вычисления времени, не превышающую 1 мс.

Большинство современных x86 материнских плат предоставляют APIC (Advanced Programmable Interrupt Controller) – улучшенный программируемый контроллер прерываний – и, как следствие, таймер APIC. Частота этого таймера равна частоте шины ЦПУ, что обеспечивает высокое разрешение (порядка 10 нс) единиц измерения времени. К тому же использование локального APIC таймера не требует в отличие от PIT или RTC обращения к порту ввода/вывода. Большинство однопроцессорных машин версии ниже Pentium 4 явно запрещают использование данного таймера, выключая его в BIOS.

Также в системах, поддерживающих ACPI (Advanced Configuration and Power Interface, усовершенствованный интерфейс конфигурации и управления питанием), представлен т.н. PM таймер (Power Management timer, таймер управления питанием), который способен (в отличие, например, от APIC таймера) предоставлять надежное время вне зависимости от изменения скорости работы процессора, связанного с использованием активного управления электропитанием.

В начале 2000-х годов компаниями Intel и Microsoft совместно был разработан таймер событий высокой точности HPET (High Precision Event Timer) [9]. Данный таймер обладает высокой частотой (не менее 10 МГц) и использует 64-битные счетчики. Зачастую он является наиболее предпочтительным высокоточным источником времени в системе.

Помимо периферийных таймеров компьютеры архитектуры x86, начиная с процессоров Pentium, имеют напроцессорный 64-битный счетчик времени TSC (Time Stamp Counter). По сравнению с остальными счетчиками он обладает такими преимуществами, как наименьшая задержка чтения и высокое разрешение. Частота обновления данного счетчика на разных процессорах может быть различна и может равняться частоте процессора либо частоте шины ЦПУ. Основные проблемы использования данного счетчика в качестве источника времени — это рассинхронизация счетчиков между ядрами процессора [10] или непостоянная частота обновления этого счетчика.

Руководство для разработчиков [11], выпущенное компанией Intel, подробно описывает различия реализации TSC в различных семействах процессоров Intel и указывает, как распознать, что процессор имеет TSC с постоянной частотой обновления. Большинство процессоров AMD имеют TSC, в силу тех или иных обстоятельств не пригодный для использования в качестве источника времени.

Перечисленные выше средства вычисления интервалов времени в архитектуре x86 дают представление о сложности решения этой задачи наиболее точно для различных процессоров. Поддержка других аппаратных архитектур (ARM, MIPS и т. д.) еще больше увеличивает её сложность. Ещё одной проблемой является нетривиальность задачи перевода «машинных» единиц времени, применяемых в использованном устройстве, в “человеческие” (например, микросекунды, необходимые протоколу ARTCP). О сложностях, сопряженных с решением данной задачи, можно получить представление ознакомившись с докладом [12], представленным в 2005 году в Оттаве (Канада) на симпозиуме, посвященном Linux.

К счастью, ядро ОС Linux предоставляет средства для решения данных задач. Чтобы иметь возможность использовать различные аппаратные счетчики и таймеры, в Linux реализована концепция “источников времени”. Согласно этой концепции, каждая аппаратная архитектура, поддерживаемая Linux, для имеющихся в ней средств реализует посредством инициализации структуры **struct clocksource** интерфейс “источника времени” и регистрирует его в операционной системе с помощью вызова **clocksource_register_khz** или **clocksource_register_hz**. Структура **struct clocksource** имеет поле **rating**, позволяющее ОС выбирать наилучший “источник времени” из доступных на данной платформе. Наилучшему “источнику времени” соответствует зарегистрированный экземпляр структуры **struct clocksource** с наибольшим значением поля **rating**. При этом значения данного поля логически интерпретируются следующим образом: 1-99 — недопустим для использования (применим только для тестирования или в процессе загрузки), 100-199 — обладает базовой функциональностью, но нежелателен для использования, 200-299 — хороший “источник времени”, 300-399 — желательный “источник времени” (достаточно быстрый и точный), 400-499 — превосходный “источник времени” (идеальный, должен использоваться, если доступен).

Выбрав лучший “источник времени” ядро может считывать значения его счетчика, вызывая функцию, предоставляемую полем **read** структуры **struct clocksource**. Эта функция возвращает значение в виде абстрактных “машинных” единиц времени, представленных типом данных **cycle_t**. Ядро может воспользоваться значениями полей **mult** и **shift** структуры **struct clocksource** и преобразовать полученное значение в наносекунды.

Для опосредованной работы с “источниками времени” и получения значений времени в “человеческих” единицах измерения ядро Linux предоставляет различные интерфейсы. Наибольший интерес среди них представляют функции **getnstimeofday** и **getrawmonotonic**. Обе функции возвращают значение времени в виде структуры **struct timespec**, состоящей из двух полей: **tv_sec**, содержащего секунды, и **tv_nsec**, содержащего наносекунды. Наиболее существенным различием между этими функциями является то, что первая в отличие от второй корректирует возвращаемое время, учитывая значения, получаемые по NTP. Отсутствие данной корректировки в **getrawmonotonic** позволяет использовать эту функцию для вычисления интервалов времени, когда соответствие значения времени, установленного на машине, значению реального времени не играет роли. Исходя из этих соображений, реализация ARTCP использует интерфейс **getrawmonotonic** для вычисления интервалов времени между двумя последовательными полученными сегментами ARTCP, содержащими полезные данные.

Первое чтение значения времени с помощью **getrawmonotonic** происходит в функции **artcp_init** при инициализации ARTCP соединения, когда сокет, отправивший SYN пакет ARTCP, получает SYN-ACK пакет ARTCP, либо сокет получает SYN пакет ARTCP (и ARTCP разрешен в системе). Последующие чтения происходят при получении любого пакета ARTCP, содержащего полезные данные. При этом если значение поля PS полученного пакета на единицу больше (по модулю 2^{32}) значения поля PS предыдущего полученного пакета, то с помощью функции **artcp_ts_diff_to_ti** вычисляется значение поля TI для пакета-уведомления.

Функция `artcp_ts_diff_to_ti` принимает два аргумента типа `struct timespec`, представляющих интервал времени, и возвращает разность между этими моментами времени в микросекундах.

Подводя итог материалу, изложенному в данной статье, можно заключить, что, произведя описанные выше изменения в исходном коде сетевой подсистемы Linux, мы получаем реализацию обработки ARTCP пакетов (прием, отправка, формирование значений полей), полностью независимую от особенностей реализации функции управления потоком и других частей протокола ARTCP. При этом реализация обладает кроссплатформенностью и использует наилучшее доступное аппаратное средство для вычисления интервалов времени, имея возможность устанавливать размерность для поля TI вплоть до наносекунд. Стоит также отметить, что реализация не конфликтует с существующей функциональностью сетевой подсистемы Linux, позволяя последней использовать TCP соединения одновременно с ARTCP и даже переключать ARTCP соединение в режим TCP.

Список литературы

1. Алексеев И.В., Соколов В.А., Чалый Д.Ю. Моделирование и анализ транспортных протоколов в информационных сетях. Ярославль: Яросл. гос. ун-т, 2004.
2. Alekseev I.V., Sokolov V.A. Compensation Mechanism for Adaptive Rate TCP // 1-St International IEEE/Popov Seminar "Internet: Technologies A and Services". October 1999. P. 68–75.
3. Алексеев И.В., Меркулов С.А., Сивов А.А. Аспекты практической реализации протокола ARTCP на ядре Linux 2.6 // Моделирование и анализ информационных систем. 2010. Том 17, № 2. С. 144–149.
4. Postel J. Transmission Control Protocol // RFC 793 (STD7). 1981.
5. Fenner B. Experimental Values in IPv4, IPv6, ICMPv4, ICMPv6, UDP, and TCP Headers // RFC4727. 2006.
6. S. Bradner, V. Paxson. IANA Allocation Guidelines For Values In the Internet Protocol and Related Headers // RFC 2780. 2000.
7. Narten T. Assigning Experimental and Testing Numbers Considered Useful // RFC 3692. 2004.
8. J. Corbet, A. Rubini, G. Kroah-Hartman. Linux Device Drivers. 3rd edition. O'Reilly, 2005.
9. IA-PC HPET (High Precision Event Timers) Specification. Rev. 1.0a. Intel Corporation, 2004.
10. AMD Technical Bulletin – TSC Dual-Core Issue & Utility Fix. Advanced Micro Devices, Inc. 2007.

11. Intel® 64 and IA-32 Architectures Software Developer's Manual. Volume 3A: System Programming Guide, Part 1. Intel Corporation. January 2011.
12. J. Stultz, N. Aravamudan, D. Hart. We Are Not Getting Any Younger: A New Approach to Time and Timers // Proceedings of the Linux Symposium. V. 1. P. 219-232. July 2005.

ARTCP Packet Structure. Features of the Formation and Processing of ARTCP Headers in Linux Network Subsystem

Sivov A.A.

Keywords: networking, transport protocol, ARTCP, time measurement, Linux kernel

The article discusses an ARTCP header structure and some practical aspects of forming the header and the calculation of header fields. The questions of precise time dispatching of the received packets are discussed. The Linux kernel interfaces for time measurement are described as well as the clock source abstraction layer and its implementation.

Сведения об авторе:

Сивов Анатолий Александрович,
Ярославский государственный университет им. П.Г. Демидова,
аспирант