

Архив свободно распространяемого программного обеспечения ftp.chg.ru: анализ 15-летнего опыта

© Л.Н. Щур, С.А. Крашаков, А.Ю. Меньшутин, В.Л. Щур, С.К. Шикота, М.В. Григорьева

Научный центр Российской академии наук в Черноголовке
dep@chg.ru

Аннотация

Приведены цели и задачи архива свободно распространяемого программного обеспечения. Обсуждаются статусные аспекты использования программного обеспечения. Обсуждается роль свободно распространяемого программного обеспечения в развитии информационных технологий. Приводятся статистические данные по использованию архива программного обеспечения, по составу программного обеспечения и его востребованности. Обсуждаются тенденции развития архивов.

1 Введение. Свободно распространяемое программное обеспечение

Архив необходим для обеспечения широкого спектра областей экономики России свободно распространяемым программным обеспечением, которое используется в персональных компьютерах, рабочих станциях, вычислительных системах и компьютерных средствах управления приборами и системами. В некотором смысле архив можно отнести к электронной библиотеке по программному обеспечению, поскольку он состоит не только из ftp сервера, но и снабжен справочной литературой, а также системами поиска программ и мониторинга его использования.

Широкое распространение персональных компьютеров в конце 1980-х годов привело к качественным изменениям в двух аспектах использования вычислительной техники. Во-первых, появилось новое качество в администрировании вычислительной техники. Администрирование перешло с уровня организации на уровень подразделений, а также на индивидуальный уровень, что значительно увеличило число системных администраторов. Во-вторых, ранее системное программное обеспечение в основном создавалось фирмами-разработчиками вычислительных систем.

В последние двадцать лет вклад независимых

разработчиков значительно вырос. В результате некоторые программные системы, созданные независимыми разработчиками, начали влиять не только на рынок программных продуктов, но и на развитие вычислительных систем в целом. Яркий и хорошо известный пример такой разработки – ядро операционной системы Linux, версия 1.0.0 которого вышла 14 марта 1994 года. Стоит отметить, что понадобилось еще много лет (не менее пяти), прежде чем стабильность работы системы достигла уровня применимости системы Linux на серверах и сетевых устройствах. До этого времени намного более популярной среди системных и сетевых администраторов была другая система – FreeBSD, надежная и устойчивая в работе¹. Именно по этим причинам архив СПО ftp.chg.ru до сих работает под управлением FreeBSD.

Обе упомянутые системы, Linux и FreeBSD, являются примерами открытого программного обеспечения. Создатель ядра Linux Linus Torvalds связывает успех его проекта в немалой степени с тем, что он лицензировал права на Linux в рамках проекта GPL.

GPL (*Открытое лицензионное соглашение GNU*) – лицензия на свободное программное обеспечение, разработанная в рамках проекта GNU в 1988 г. Свободно распространяемое в рамках лицензии GPL программное обеспечение играет большую роль в развитии информационных технологий. Оно является платформой для построения новых систем при скромном бюджете, но требуя при этом наличие квалифицированных специалистов. Его роль, таким образом, особо важна в развивающихся странах с относительно слабой экономикой, но высоким уровнем образования населения, например, в таких странах, как Бразилия, Россия, Индия и Китай.

2 Архив ftp.chg.ru – анализ развития

Архив программного обеспечения ftp.chg.ru – это проект, развиваемый в Научном центре РАН в Черноголовке, в Отделе прикладных сетевых иссле-

Труды 12^й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010

¹ Например, один из маршрутизаторов оптоволоконной опорной сети Chg-FREENet проработал на версии FreeBSD 1.1.5.1 более пяти лет с августа 1994 года до февраля 2000 года, причем не только без обновлений, но и без единой перезагрузки. Он был демонтирован в связи с переводом сети на другую технологию

дований. Основная идея открытого программного обеспечения (ПО) состоит в том, что каждый участник может внести свой вклад в развитие этого ПО. Помощь могут оказывать не только профессиональные программисты, занимающиеся разработкой ОС или прикладного ПО, но и пользователи ПО. Для успешного развития требуется помощь как пользователей, предоставляющих отчеты об ошибках, так и ресурс-провайдеров. Поскольку большинство открытых проектов ограничено в финансовых ресурсах, очень часто для хостинга таких проектов используются серверы учреждений, где работают авторы проекта. Однако для распространения действительно крупных и популярных проектов требуется помощь сообщества. Одним из таких серверов, предоставляющим услуги хостинга на безвозмездных условиях различным проектам, является сервер ftp.chg.ru.

2.1 История развития сети Научного центра РАН в Черноголовке

История архива ftp.chg.ru неразрывно связана с историей развития сети Научного центра Российской академии наук в Черноголовке (НЦЧ). При начале работ по созданию сети в 1989 г. ставилась задача обеспечения интерактивного доступа к ресурсам интернета с рабочего места каждого научного сотрудника. Сегодня решение такой задачи обеспечивает армия интернет сервис-провайдеров, а 20 лет назад такая постановка казалась нереальной даже в академической среде, которая всегда имела много контактов как внутри страны, так и по всему миру. Основным средством общения в то время были телекс и факс. В то время еще не было мировой паутины веб, да и интернета тоже не было. Глобальная сеть виделась как гетерогенное образование, объединяющее сети фирм Digital (DECnet), IBM (BITNET), Netware (Novell) и сети на протоколах X25 и IP. Начали создаваться межсетевые шлюзы. Эта работа велась в научных учреждениях всего мира.

В 1989 году в НЦЧ РАН была построена первая опорная сеть на основе протокола X25. Она использовалась для отправки и получения электронных писем через узел КИАЭ. В 1990 году был организован доступ к узлу BITNET в ИОХ РАН. В марте 1992 года были созданы IP-узлы сети FREEnet в ИТФ им. Л.Д. Ландау РАН в Москве и Черноголовке. В 1993 году была построена опорная сеть на выделенных медных парах в НЦЧ РАН и выделенный канал сети FREEnet. С этого момента началась интерактивная работа сотрудников институтов Черноголовки в интернете.

Это было время, когда веб-сайты были текстовыми, первый графический браузер Mosaic был создан в NCSA (National Center for Supercomputing Applications, Urbana-Champaign, Illinois) в 1993 году. В том же 1993 году под руководством вице-президента США А. Гора была разработана национальная программа развития интернета. Именно эта

программа придала развитию информационных технологий взрывной характер.

В 1994 году была начата работа по переводу опорной сети НЦЧ РАН на оптоволоконно, и первые сегменты были сданы в работу в августе. По имеющейся информации, это была первая оптоволоконная сеть в Академии наук, да и, возможно, во всем научном секторе России.

В настоящее время сеть НЦЧ РАН является независимой академической сетью и имеет название ChANT (Chernogolovka Academic Network). Она входит в ассоциацию научно-образовательных сетей России e-Arena. Опорная оптоволоконная сеть имеет топологию двух колец, на скоростях 1 гигабит в секунду и 10 гигабит в секунду. Внешний оптоволоконный канал создан совместно НЦЧ РАН и ОАО Ростелеком. Сеть ChANT в настоящее время использует 155 мбит в секунду для подключения к точке обмена научно-образовательных сетей NAP (Network Access Point – коннеktivность с научными сетями Евросоюза и Северной Америки), к точке обмена российских интернет-провайдеров MSK-IX (Moscow Internet Exchange) и для выхода в сети общего пользования.

2.2 История архива программного обеспечения ftp.chg.ru

Развитие сети Научного центра требовало соответствующего программного обеспечения. В первую очередь требовались системное и сетевое ПО для создания инфраструктуры сети. Далее, необходимо было ПО для работы пользователей – системы подготовки текстов статей, рисунков, графиков и т. п. При ограниченности пропускной способности каналов связи, да и различного рода других ограничений известного характера появилась необходимость в создании в структуре сети НЦЧ РАН хранилища для программного обеспечения, или ftp-архива. Такой архив был создан в ИТФ им. Л.Д. Ландау РАН, где сеть DECnet работала с 1989 года и была соединена шлюзом с IP-узлом на основе пакета KA9Q в 1991 году. Этот узел обеспечивал обмен e-mail с остальным миром. С 1993 года связь с внешним миром стала интерактивной. Внешний канал к концу 1993 года не превышал 64 Кбит в секунду.

Краткая история объема архива и скорости доступа к нему такова:

- 1995 – перемещение архива на отдельный сервер с именем ftp.chg.ru, объем архива 1 – 2 Гб;
- 1997, июнь – запуск 2-х мегабитного канала Черноголовка – Москва;
- 1997, июль – август – активное наполнение ftp-архива и начало активного использования архива пользователями за пределами Черноголовки;
- 1997, декабрь – 1-е место в Top100 России, объем архива ~ 40 Гб;
- 2000, ноябрь – перенос архива в Москву (ЮМОС) и подключение на скорости 10 Mbps, объем архива ~ 350 Гб;

- 2001, август – перевод на 100 Mbps, объем архива ~ 495 Гб;
- 2003 г. – начало работы по протоколу IPv6;
- 2004 г. – объем архива перешагнул рубеж 2 Тбайта;
- 2008 г. – объем архива 32 Терабайта.

Архив широко известен в мире. С 1997 года и по настоящее время архив ftp.chg.ru занимает места с 3-го по 5-е в мировой иерархии архивов свободно распространяемого программного обеспечения. К сожалению, на сегодняшний день оценить точное место нашего архива в мировой таблице о рангах достаточно сложно. Так, последняя доступная статистика на сайте <http://www.ftp-sites.org/> датирована 2003 г. Имеющиеся у авторов собственные наработки по мониторингу мировых ftp-серверов тоже дают противоречивые результаты т. к. получение полного листинга файлов из большинства крупных серверов запрещено вследствие большого объема генерируемой информации.

2.2 Технологии

Краткое описание системы:

- Аппаратная часть архива состоит из двух серверов с RAID массивами, реализующими функции хранилища. Каждый сервер основан на 2-х процессорных платах Intel и 2-х ядерных процессорах семейства Xeon. Объем оперативной памяти каждого сервера – 4 Гб.
- Программная часть базируется на операционной системе свободного доступа FreeBSD. Текущие версии – 6.1 и 8.0.
- Накопление массива построено на регулярном обновлении программного обеспечения по заранее составленному списку.
- Значительная часть ПО обновляется с оригинальных мастер-серверов по специальной подписке.
- Архив доступен как по протоколу IPv4, так и по протоколу IPv6. Обновление информации происходит с преимущественным доступом по протоколу IPv6. Так, средняя скорость входящего трафика по протоколу IPv6 – 700 Кбайт/с, а максимальная – до 2 Мбайт/с. При этом скорость входящего трафика по протоколу IPv4 – тоже около 700 Кбайт/с.

2.3 Программный состав архива

Программное обеспечение архива ftp.chg.ru можно разбить на такие группы:

- Операционные системы семейства: FreeBSD, FreeSBIE, DragonFlyBSD, NetBSD, OpenBSD;
- Базы данных: PostgreSQL, MySQL, mSQL, SAPdb, FreeTDS;
- Стандарты интернета: RFC, Standards, Internet Drafts, Usenet FAQ's, Russian Doc's;
- Операционные системы семейства Linux: ALTLinux, ASPLinux, BLin, Conectiva, Corel Linux, Debian, Debian-CD, Fedora Linux, GamesKnoppix,

Gentoo Linux, Gnppix , kernels, Mandrakelinux, RedHat, Slackware, SOT Linux, SunSite collection, SuSE, Ubuntu Linux;

- Мультимедийные системы: ImageMagick, MultimediaMultimedia, Geomview, Portable Network Graphics, UC Berkeley Plateau Multimedia Project, Other Graphics;
- Сетевые системы: IPv6 Stacks, cmu-snmp, Netperf, Nocol, MBone, Multi Router Traffic Grapher, Ttcp, zebra;
- Программы для Windows: Antiviruses, FreeDOS, Cygwin, Simtel.NET, WinSite;
- Системы и языки программирования: CPAN – Comprehensive Perl Archive Network, EGCS, GNU Science Library, Java, Lisp, Netlib, Parallel Programming Libraries & Tools, Python, Smalltalk, Tcl;
- Системы для науки и образования: Computational Chemistry, Mathematics, Molecular biology, Statlib;
- Системы безопасности: DFN-CERT;
- Системы обработки текстов: CTAN, TeX live, OpenOffice.org, OpenOffice-RU;
- Надстройки для систем типа Unix: Benchmarks, Gnu directory, EGCS, Gimp, GNOME, KDE, kernel.org mirror, Freeware for Solaris, Pine, Samba, Berkeley Sendmail, X11;
- Браузеры, Веб-серверы и т.п.: Analog, Apache, Lynx, Netscape, Opera, Mozilla, Squid.

2.4 Пользователи

Число одновременно работающих пользователей – более 5 тысяч, из них большая часть по протоколу http. При этом максимальное число пользователей ограничено искусственно в конфигурации сервера, т. к. большое число одновременно открытых соединений, время жизни которых часто превышает десятки минут, приводит к быстрому исчерпанию оперативной памяти и вызывает нестабильную работу системы.

2.5 География

Свободно распространяемое ПО является практически всегда универсальным и пригодным для использования пользователями со всего мира. Так, любой дистрибутив Linux или FreeBSD содержит в себе переводы на большинство языков. Поэтому пользователь может скачать интересующий его файл с любого зеркала, в том числе и с нашего архива. За все время работы архива мы сохраняем подробную статистику запросов. Анализ такой статистики может представлять интерес для большого круга лиц. Например, можно ответить на извечный вопрос, какой дистрибутив Линукса наиболее популярен в данный момент. Определенный интерес представляет и статистика географии запросов.

География запросов включает в себя практически все домены мировой сети. Интересно, что доля

запросов из отдельных регионов может значительно флуктуировать. Например, в октябре 2007 года 20 процентов исходящего трафика было в зону Франции. На текущий момент основная доля запросов приходится на Россию (2 млн. за месяц, 2,5 Тб – февраль 2010). На втором месте, по числу запросов, находится Китай (80 тыс. запросов, 16 Гб трафика). Украина занимает 3-е место по числу запросов (70 тыс.) и 2-е по объему трафика (120 Гб).

Предположительно, большое число запросов и малый объем трафика из Китая связаны с тем, что данные запросы являются запросами на автоматическое обновление ПО в различных дистрибутивах, зеркалом которых является ftp.chg.ru. Такие обновления обычно не велики по размеру, поэтому общий трафик от них невелик.

В некоторые часы заметна активность Австралии, что может быть связано с особенностями позиции локального светового дня.

2.6 Статистика

Для доступа удобства работы со статистикой нами было разработано специальное ПО, которое анализирует лог-файлы и сохраняет данную информацию в базу данных. Веб-интерфейс с этой базы данных содержит не только данные о количестве запросов к тому или иному ресурсу, но также описания ресурсов. В частности, содержится информация о разных Линукс-дистрибутивах и т. п. Веб-интерфейс доступен по адресу <http://archive.chg.ru/>.

Запросы могут быть сделаны за определенный день или любой другой интервал. Запросы могут быть сделаны по отдельным пакетам, а также по ветвям дерева файловой системы.

2.7 Процедура ротации лог-файлов

Анализ статистики требует определенных усилий, т. к. некоторые закачки продолжают более одного дня. Чтобы разбить общий лог-файл на несколько файлов, содержащих данные за определенный день, достаточно попросту перезапустить веб или ftp-сервер, предварительно скопировав старый лог-файл в другое место. Сервер FTP работает через промежуточный демон inetd, т. е. для каждого нового соединения запускается своя копия программы. При этом при перемещении старого лог-файла в новое место новый лог-файл будет создан автоматически при запуске новой копии программы. Поэтому никаких особых действий для форсирования создания нового лог-файла для сервера протокола FTP не требуется. Аналогично серверу ftp работает и сервер rsync.

Для ротации лог-файлов http-сервера и подготовки их к дальнейшему анализу написана программа на языке perl – `rotate.pl`. Алгоритм действий, выполняемых данной программой, таков:

- текущий лог-файл – `/usr1/www/log/httpd-access-ftp.log`, временная копия – `/usr1/www/log/httpd-access-ftp.log.old`;
- если временная копия используется каким-

нибудь процессом (демоном `httpd`) – завершить работу программы;

- разбить временный файл на части по дням;
- удалить временный файл и переименовать текущий лог файл во временный;
- послать сигнал (при необходимости) о необходимости создания нового лог-файла процессам-демонам (`Apache`).

Разбивка файла по дням осуществляется на уровне отдельных записей в файле на основе имеющейся метки времени. Запись в лог-файле, которая относится к определенному промежутку времени, добавляется в конец отдельного лог-файла, хранящего информацию о запросах за определенный день по определенному протоколу.

2.8 Анализ статистики

Разбитые по дням лог-файлы подвергаются дальнейшему анализу. Так, производится анализ с целью выявления наиболее популярных ресурсов, анализ количества запросов к тому или иному ресурсу и т. п. Для корректной обработки статистики по описанию ресурсов разработана специальная программа. Поскольку ресурс (например, Линукс определенной версии) не обязательно находится в одном каталоге, а может находиться в разных местах файловой системы, описание ресурса ведется путем перечисления всех каталогов (с использованием регулярных выражений), принадлежащих данному ресурсу. Для того чтобы избежать частой необходимости исправления/дописывания описаний имеющихся ресурсов, регулярные выражения могут использоваться и в именах ресурсов. Таким образом, одно описание может охватывать сразу все версии того или иного программного продукта.

Анализ статистики запросов показывает, что наиболее популярными файлами являются файлы `*.deb`. Очевидно, данный факт связан с ростом популярности ОС на базе Debian Linux, в т. ч. Ubuntu. Однако по количеству трафика лидируют файлы `*.iso` с дистрибутивами Linux, а также пакет OpenOffice версии для Windows.

2.9 Балансировка нагрузки

Серверы файлового архива ftp.chg.ru предоставляют доступ пользователям по протоколам HTTP, FTP и RSYNC. Балансировку загрузки по протоколу HTTP достаточно несложно реализовать, используя возможности протокола по переадресации запросов на другой ресурс. К сожалению, указанная возможность отсутствует в протоколах FTP и RSYNC, поэтому данный вид балансировки загрузки неприменим к этим протоколам.

Сказанное означает, что единственным возможным вариантом, призванным обеспечить балансировку загрузки между двумя серверами ftp.chg.ru и ftp.chant.ru, расположенными соответственно в Москве и Черноголовке, является механизм балансировки на уровне сервера DNS. Сервер ftp.chant.ru имеет ограниченную связность с внешними сетями.

При этом к сети НЦЧ РАН он подключен на скорости 1Gbit/s. Поэтому естественным выбором является следующая схема организации работы: все запросы на сервер ftp.chg.ru не из научной сети НЦЧ РАН должны направляться на сервер, расположенный в Москве. Все запросы, исходящие из сети НЦЧ РАН, на сервер в Черноголовке.

Данная схема была реализована посредством специальной конфигурации серверов DNS ns1.chant.ru ns2.chant.ru и ns3.chant.ru. При необходимости часть запросов с сервера ftp.chg.ru может быть перенаправлена на сервер, расположенный в Черноголовке, путем внесения нескольких записей в таблицу сервера DNS, указывающих на сервер номер 1, и нескольких записей, указывающих на сервер 2. Варьируя количество записей и их соотношение, можно добиться контроля количества запросов к определенному серверу.

Единственным недостатком данной схемы является то, что время действия записи в таблице DNS-сервера должно быть достаточно велико и составлять несколько часов. Если же данный параметр будет значительно уменьшен для обеспечения возможности более быстрого изменения схемы перераспределения запросов, то это может вызвать повышенную нагрузку на серверы службы DNS.

3. Обсуждение. Перспективы

Свободное ПО интенсивно развивается. Растет и

его объем. Например, объем операционной системы FreeBSD вырос с 15 Гб в 1998 году до 1 Тб в настоящее время. Рост объема архива связан с определенными затратами как финансового, так и ресурсного характера.

Необходимо также отметить парадокс в разрыве отношения к свободному ПО со стороны административных структур и пользователей. Если первый падает, то второй, напротив, растет. Есть надежда, что с созданием Российской ассоциации свободного программного обеспечения (РАСПО) использование свободного ПО в России получит новый толчок.

Мы надеемся на интерес спонсоров к нашему общественному проекту архива свободно распространяемого ПО.

Archive of public domain software ftp.chg.ru: 15 years of development

L.N. Shchur, S.A. Krashakov, A.Yu. Menshutin,
V.L. Shchur, S.K. Shikota, M.V. Grigorieva

We present project of open software archive. Practical aspects of open software usage discussed. Particular interest paid for the influence of open software development on the progress in information technologies. Statistics of archive usage discussed, as well as the future development of archive.