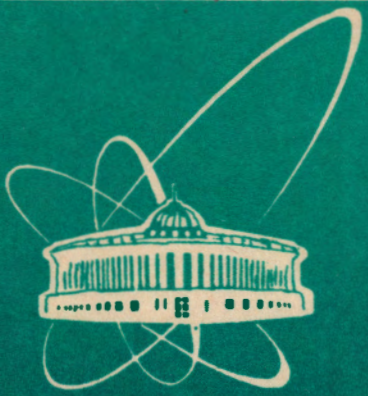

сообщения
объединенного
института
ядерных
исследований
дубна

Р11-93-221

А.А.Вовенко

НЕКОТОРЫЕ ОСОБЕННОСТИ ПРИМЕНЕНИЯ НАКОПИТЕЛЕЙ ИНФОРМАЦИИ ТИПА EXABYTE В СИСТЕМАХ РЕАЛЬНОГО ВРЕМЕНИ

1993

1 Введение

Появление в последнее время цифровых магнитофонов с записью информации на видеокассеты означает значительный прогресс в системах хранения информации на магнитных носителях. На одну кассету размещается до 5 Гбайт информации, что эквивалентно 50 – 200 обычным магнитным лентам. При этом время позиционирования ленты в произвольно указанное место не превышает полутора минут. Но использование технических средств нового поколения всегда влечет преодоление новых трудностей. Во-первых, это необходимость обеспечения целостности огромного количества информации, находящейся на кассете, в том числе в ситуациях, когда возможны ошибки в работе персонала, программах, сбои аппаратуры. Во-вторых,это проблемы эффективного управления аппаратурой для достижения хороших показателей быстродействия.

При применении таких устройств в системах реального времени жизненно важно обеспечить успешную запись информации с первой попытки, поскольку современные эксперименты требуют значительных затрат времени и весьма дорогостоящи и их повторное проведение всегда проблематично.

2 Надежность

Под надежностью системы хранения информации следует понимать вероятность записи и последующего извлечения достоверных сведений с ленты после периода хранения и транспортировки ленты. Она складывается из надежности аппаратных средств (лент, накопителей) и устойчивости программ к различного рода сбоям и нештатным ситуациям. Последнее особенно актуально для носителей с очень большой емкостью, типа Exabyte, поскольку одно неправильное действие может привести к потере всех ранее записанных данных на ленте.

2.1 Надежность носителя

Встроенный аппаратный контроль записи (чтение с жестким контролем уровня остаточной намагниченности ленты), запись 400 байт контрольных сумм на каждые 1024 байта данных и использование алгоритмов, восстанавливающих множественные ошибки, гарантируют надежное долговременное хранение данных на ленте и возможность их дальнейшего успешного извлечения.

Эффективный механизм контроля записи привел к появлению на Западе в вычислительных центрах, где используется большое число кассет, практики, когда для записи информации используются обычные видеокассеты, не сертифицированные для цифровой записи. При этом в партии из 100 кассет 2 или 3 оказываются негодными (это выясняется в течение первых пяти минут записи), а остальные успешно работают. Следует оговориться, что тем не менее выбираются строго определенные модели кассет фирмы Sony.

2.2 Взаимозаменяемость

Фирма Exabyte провозглашает полную взаимозаменяемость накопителей и совместимость от младших моделей к старшим. Имея ограниченный опыт переноса кассет с одного накопителя на другой, автор тем не менее может констатировать, что качество чтения кассеты на другом стримере значительно хуже, чем на том, где она была записана. Проявляется это в том, что на порядок чаще включается механизм коррекции ошибок, иногда производятся повторные чтения блоков, а это приводит к значительным паузам (10–20 секунд) при чтении, и иногда могут даже возникать отказы чтения.

К сожалению, сейчас не представляется возможным набрать какую-либо статистику по взаимозаменяемости накопителей ввиду их малого количества, труднодоступности и т. д. Автор будет очень признателен за любую информацию по этому поводу.

2.3 Безопасность программных средств

Основная задача для обеспечения безопасной работы программы заключается в предотвращении случайной записи при неправильно спозиционированной ленте. Дело в том, что после выполнения операции записи внутренний контроллер стримера автоматически записывает специальную метку конца ленты. При операции чтения невозможно спозиционировать ленту далее этой метки конца ленты. Метка распознается при всех операциях движения ленты, включая быстрое позиционирование. Таким образом, если в начале ленты случайно или преднамеренно перезаписать хотя бы 1 блок, то все содержимое ленты становится недоступным, хотя физически на ленте будет затерто очень небольшое количество информации. Чтение данных с таких лент хотя, в принципе, и возможно , но является нетривиальной операцией и сопровождается потерей информации с того участка ленты, где была произведена перезапись (метка конца ленты занимает на ленте пространство, эквивалентное нескольким Мбайт данных).

Таким образом, главными задачами для обеспечения сохранности данных являются задачи контроля за положением ленты, предотвращения одновременного доступа из нескольких разных программ и предотвращения нежелательных операций записи.

Существующая система on-line приема информации ориентирована на режим работы с ускорителем У-70, в котором цикл ускорения занимает 8 секунд, а время вывода пучка до 1,5 секунд . Соответственно, в течение времени вывода пучка программа аккумулирует информацию, а в промежуток между сбросами осуществляет сборку, контроль и запись накопленной информации. Система функционирует на IBM/PC под управлением MS-DOS, которая не имеет, к сожалению, средств работы с лентами.

Для контроля за положением ленты используются следующие методы:

- Блокировка органов управления на передней панели устройства при активном Run. Run – это сеанс набора данных, при одинаковом режиме работы ускорителя и детектора. Из практических соображений максимальный объем информации в одном Run ограничен 32 Мбайтами.
- Автоматическое позиционирование к метке конца ленты при начале любого Run. Для этого используется специальный вариант SCSI команды Space Blocks.
- Выполнение контроля положения ленты перед началом очередного цикла записи данных (каждые 8 секунд). В связи с отсутствием в MS-DOS механизмов контроля доступа к внешним устройствам приходится самостоятельно принимать меры по защите от одновременного обращения из нескольких программ. Эти же меры должны защищать содержимое ленты при сбоях подсистемы SCSI, при восстановлении после которых следует команда SCSI reset и вызванная ей нежелательная перемотка ленты на начало, которую,к сожалению, невозможно отключить.

Перед началом записи массива информации производится запрос положения ленты. Накопители Exabyte всегда знают положение ленты и сообщают абсолютный номер блока, начиная с начала ленты. Полученное таким образом число сравнивается с абсолютным номером блока в предыдущем цикле. Если текущий номер блока меньше, чем в предыдущем цикле,то это означает, что произошло

несанкционированное движение ленты, запрос на запись отвергается и выдается сообщение об ошибке.

Этот алгоритм также защищает ленту от ошибок в программе сбора данных и не дает вести запись на ленту, если Run не был начат (и лента не была спозиционирована к концу данных).

2.4 Алгоритм прекращения операций записи

В конце Run производится сброс всех буферов, стоящих в очереди на запись, на ленту и запись метки конца ленты. Накопители Exabyte устроены так, что команда записи метки конца файла вызывает автоматическую запись всех блоков данных, стоящих в очереди на запись. Таким образом, после выполнения команды записи метки конца файла все данные и сама метка конца файла оказываются фактически записанными на ленте. Но при этом Exabyte не выполняет запись метки конца ленты, т.к. после записи потребовалась бы операция позиционирования назад, в положение перед меткой конца ленты, а это заняло бы слишком много времени (около 30 секунд).

Таким образом, после операции записи метки конца файла на ленте не записывается автоматически метка конца ленты. Если оставить устройство в таком состоянии, то метка не будет записана вообще, если далее будет выключено питание или подана команда SCSI Reset.

Экспериментально установлено, что единственно надежным способом провоцирования выполнения записи метки конца ленты является выполнение команды движения назад. Это вносит задержку в десятки секунд в конце Run, но гарантирует фактическое выполнение записи всех данных и служебных полей устройством Exabyte. Более подробно о необходимости метки конца ленты см. раздел 2.5 "Обработка исключительных ситуаций".

2.5 Обработка исключительных ситуаций

Чаще всего возникают следующие исключительные ситуации: зависание компьютера с программой сбора данных и зависание шины SCSI. Их причинами могут являться кратковременное пропадание питания в сети, электромагнитные наводки, а также ошибки в программном обеспечении.

При сбое компьютера желательно избежать выдачи команды SCSI Reset при перезагрузке программного обеспечения. Эта задача решается соответствующей настройкой драйверов SCSI или модификацией их

двоичного кода. В этом случае сохраняется положение ленты в конце данных, сохраняется соответствие состояния ленты и накопителя Exabyte. Если смириться с отсутствием метки конца предыдущего Run на ленте, то можно сразу продолжить запись на ленту.

При зависании SCSI или после сбоя питания необходима выдача команды SCSI Reset. По этой команде производится полная инициализация устройства и все команды, стоявшие в очереди на выполнение внутри Exabyte, игнорируются. Как правило, в очереди на выполнение стоит операция записи метки конца ленты. Это длительная операция (десятки секунд), и она производится фактически только тогда, когда Exabyte обнаруживает переход от команд записи к командам чтения/позиционирования ленты. В нашей версии накопителей и документации есть расхождение между описанием в книге и фактическим поведением устройств. Документация утверждает, что некоторые другие команды также вызывают запись метки конца ленты, но фактически этого не происходит. По всей видимости, изменения были внесены в аппаратуру в последний момент для повышения ее быстродействия и еще не отражены в документации.

Таким образом, при сбое SCSI или пропадании питания лента оказывается в состоянии без метки конца ленты. Отсутствующая метка конца ленты диагностируется при попытке чтения/записи/позиционирования внутренним контроллером Exabyte как брак носителя информации (Media Error) и приводит к отказу от операции записи. Это порождает ряд проблем, даже если было точно известно положение ленты к моменту сбоя.

Дело в том, что контроллер Exabyte не позволяет начать операцию записи в произвольном месте на ленте. Запись разрешена только в следующих местах:

- в начале ленты,
- у метки конца ленты,
- после метки конца файла,
- перед меткой конца файла.

Это означает, что дозапись на такую ленту невозможна. Единственным выходом является позиционирование ленты назад, к ближайшей метке конца файла, и возобновление записи с этого места. В нашей версии программного обеспечения метка конца файла пишется только между Run, размер которых составляет до 32 Мбайт. Выполнение этой операции воз-

можно либо вручную при помощи интерактивной программы работы с Exabyte, либо автоматически, при помощи специально написанной утилиты анализа и ремонта ленты. В любом случае данные последнего незавершенного Run теряются.

В принципе, их можно было бы сохранить, если скопировать их на диск, восстановить структуру ленты, а затем вернуть эти данные на ленту и добавить метку конца Run. Для этого требуется только свободное дисковое пространство в 32 Мбайта и время для копирования. Простой расчет показывает, что время, требуемое для копирования данных на наш диск и снова на ленту, сопоставимо с временем набора новых данных. Поэтому возможность восстановления этих данных хотя и желательна, но не является крайне необходимой. А отсутствие свободного дискового пространства отодвигает эту задачу на третий план.

3 Быстродействие

Накопители Exabyte-8500 являются буферизированными устройствами и обладают собственной буферной памятью размером в 1 Мбайт. При операциях чтения эта память используется для размещения заранее прочитанных данных, при операциях записи блоки данных также помещаются в буферную память и сама операция записи начинается, когда объем данных в буфере превышает порог (512K по умолчанию). Это производится для уменьшения времени реакции на команды чтения/записи блока. При такой организации накопителя затраты времени программой на чтение или запись блока сводятся только к затратам на пересылку данных из компьютера в буферную память или наоборот. Последнее утверждение справедливо для случаев, когда темп обмена данными не превышает предельную скорость физического чтения/записи ленты. Для устройства Exabyte-8500 это 500 Кбайт/с.

Это позволяет простыми средствами строить системы с удовлетворительными характеристиками по быстродействию. Буферная память освобождает программу от ожидания выполнения команд механическим приводом в основных режимах (продолжительная запись или чтение без операций поиска или перемотки).

Документация на Exabyte [2] и контроллер Adaptec-1542 [3] провозглашают следующие данные по быстродействию:

Средний темп обмена данными при непрерывном движении ленты	500 Кбайт/с
Пиковый темп обмена данными по шине SCSI при асинхронной передаче данных	1.5 Мбайт/с
Пиковый темп обмена данными по шине SCSI при синхронной передаче данных	4.0 Мбайт/с

Происхождение этих цифр следующее: средний темп обмена данными при непрерывном движении ленты (500 Кбайт/с) – это тот темп, в котором данные считываются магнитной головкой и обрабатываются системами декодирования, вычисления контрольных сумм и т. д. Как следует из названия этой величины, такая скорость чтения или записи может достигаться в среднем, в течение большого промежутка времени, при условии, конечно, что процесс идет без сбоев, подсистемы коррекции ошибок не включаются и сторону, принимающую или поставляющую данные для Exabyte, никогда не приходится ожидать.

Пиковый темп обмена данными по шине SCSI указан, по всей видимости, исходя из схемотехнических решений контроллера DMA на PC и микропрограммного автомата, управляющего циклами шины SCSI. Эти цифры характеризуют темп обмена байтами данных *в процессе* передачи блока. Это очень важная характеристика быстродействия системы, но, к сожалению, не полная.

Ни одна из цифр, указанных фирмами, не дает ответа на следующий вопрос: а сколько времени будет занимать запись одного блока данных, при условии, что для него достаточно места в буферной памяти. Дело в том, что для пересылки блока данных оба участвующих в пересылке контроллера должны произвести определенные действия по управлению своей аппаратурой и по осуществлению протокола SCSI. Этот протокол достаточно сложен и на практике всегда реализуется посредством микропроцессора с зашитой в него программой. В контроллеры такого типа ставятся недорогие однокристальные процессоры, работающие, естественно, медленнее, чем центральный процессор компьютера. Таким образом, требуется время на выполнение какого-то фрагмента программы в микропроцессорах контроллеров SCSI и Exabyte. Полное время записи блока складывается, таким образом, из времени проведения подготовительных операций и собственно из времени пересылки данных. Естественно ожидать, что для блоков большой длины время будет приближаться к теоретическому пределу, поскольку там основной вклад будет вносить процесс пересылки данных. А для блоков маленького размера следует ожидать значительно более низкое быстродействие.

На детальное исследование вопросов скорости работы SCSI меня

Таблица 1: Полные времена выполнения отдельных команд

Операция	Асинх. реж.	Синх. реж
Тест готовности устройства	5.6 мс	
Чтение диапазона длин блоков	8 мс	
Inquiry	52 мс	
Mode Sense	13 мс	
Mode Select	9 мс	
Запрет извлечения кассеты	5.3 мс	5.3 мс
Request Sense	18 мс	8 мс
Чтение блока в 160 байт	8.4 мс	8.4 мс
Чтение блока в 16 Кбайт	20 мс	11 мс
Чтение блока в 160 байт с Request Sense	27 мс	
Чтение блока в 16 Кбайт с Request Sense	38 мс	
Попытка записи на защищенную ленту	6.6 мс	
Запись блока в 160 байт	7.5 мс	7.5 мс
Запись блока в 16 Кбайт	18 мс	12.4 мс
Запись блока в 32 Кбайт	28.5 мс	17.2 мс
Пропуск метки конца файла	9.0 мс	9.2 мс
Пропуск блока из 160 байт	9.0 мс	8.4 мс
Пропуск блока из 16 Кбайт	11.8 мс	11.8 мс

толкнуло следующее неприятное открытие: простейшая программа копирования лент с одного накопителя Exabyte на другой не может достичь темпа в 500 Кбайт/с. (В моей программе происходила двойная передача данных: из Exabyte в компьютер, а из него в другой Exabyte). При двойных пересылках нагрузка на SCSI должна была составить 1 Мбайт/с, что при пиковой способности в 4 Мбайт/с (в синхронном режиме) не должно являться ограничивающим фактором. Однако пробное копирование ленты заняло не два часа, а все шесть, что побудило меня проделать серию замеров времен выполнения отдельных команд.

Времена, приведенные в табл.1, – это полные времена, необходимые на выполнение команд, для пары Exabyte-8500 — Adaptec-1542, включая времена, потраченные центральным процессором в драйвере Aspi4Dos при инициализации команды.

Беглый взгляд на таблицу показывает, что для выполнения любой команды требуется время порядка 8 миллисекунд. Это значительное по по-

нятиям современной технологии время. Для сравнения скажем, что время позиционирования головки с одной дорожки на соседнюю в современных дисках составляет 2 – 3 миллисекунды. И это для выполнения команды механическим приводом. А у нас тратится в три раза большее время только для выполнения одной команды, не требующей ни механических операций, ни пересылок больших массивов данных.

Сравнительный анализ времен записи блоков данных для разных длин блоков позволяет оценить действительный темп передачи данных и время, затраченное контроллерами на организацию выполнения команды. В синхронном режиме, например, 7,5 мс требуется на подготовительные операции, после чего данные начинают передаваться в темпе 3–4 Кбайт/мс. Мы видим, что если лента записана блоками по 16Кбайт, то менее половины времени используется для передачи данных, а остальное время центральный процессор и канал DMA свободны и простаивают. Таким образом, чтобы эффективно задействовать ресурсы подсистемы SCSI, надо значительно увеличить объем данных, приходящихся на 1 команду. Это можно сделать двумя способами: или существенно увеличив размер блока на ленте (например до 160 Кбайт), или передавая/принимая по несколько блоков за одну команду. Оба эти решения требуют наличия значительного размера буферной памяти, для помещения в него большого блока данных, но первое решение потребует всегда иметь буфер большого размера при работе с записанной таким образом лентой , в то время как во втором случае значительное количество памяти требуется только в тот момент, когда необходимо достижение высокого темпа работы, а при нехватке памяти можно перейти на чтение ленты по 1 блоку за раз. Но этот способ работы возможен только с блоками фиксированной длины.

Также следует обратить внимание на команду Request Sense. Трудно придумать объяснение, почему в асинхронном режиме на передачу 14 байт данных тратится на 10 миллисекунд больше, чем в синхронном, но и в том и в другом случае эти времена настолько велики, что крайне нежелательно пользоваться этой командой после каждого блока данных. Это делает нежелательным использование формата блоков переменной длины, как бы удобно он не подходил к реальной структуре записываемых данных, поскольку определение фактической длины делается при помощи команды Request Sense, что удваивает время, затрачиваемое на чтение каждого блока.

Таким образом, следует сделать вывод, что для обеспечения эффективной работы накопителей типа Exabyte необходимо, во-первых, записывать данные блоками фиксированной длины, во-вторых, записывать

информацию возможно большими порциями за каждую команду, что требует введения еще одного уровня буферизации, в дополнение к имеющемуся в них аппаратно.

Литература

[1] American National Standard for information systems
SMALL COMPUTER SYSTEM INTERFACE – 2 (SCSI-2)
March 9, 1990
Copies of this document may be purchased from: Global Engineering Documents, 2805 McGaw, Irvine, CA 92714 Please refer to document X3.131-198X.
BBS phone 316-636-8700

[2] EXB-8500 8mm Cartridge Tape Subsystem
User's Manual
EXABYTE Corporation 1685 38th Street
Boulder, Colorado 80301
(303)442-4333

[3] AHA–1540A/1542A USER'S MANUAL
Adaptec, Inc. 691 South Milpitas Blvd., Milpitas, CA 95035.

Рукопись поступила в издательский отдел
18 июня 1993 года.