
ГОСУДАРСТВЕННЫЙ НАУЧНЫЙ ЦЕНТР РОССИЙСКОЙ ФЕДЕРАЦИИ
**ИНСТИТУТ ФИЗИКИ ВЫСОКИХ ЭНЕРГИЙ**

ИФВЭ 2013-22
ОЭФ

С.В. Донсков, А.В. Инякин, Ю.Д. Карпеков, В.Д. Матвеев,
В.Ф. Образцов, В.И. Романовский, В.А. Сенько, М.М. Солдатов,
А.П. Филин, Н.А. Шаланда, В.И. Якимчук
*Институт физики высоких энергий, Протвино*
А.И. Макаров, А.А. Худяков
*Институт ядерных исследований РАН, Москва*

# Система сбора данных эксперимента OKA

Направлено в *ПТЭ*

Протвино 2013

УДК 519.25.256 М–24

**Аннотация**

Донсков С.В. и др. Система сбора данных эксперимента ОКА: Препринт ИФВЭ 2013-22. – Протвино, 2013. – 28 с., 8 рис., 3 табл., библиогр.: 34.

В данной работе описана система сбора данных эксперимента по изучению редких распадов каонов ОКА на ускорителе У-70. Основой блокируемой системы сбора данных послужила система МИСС (Многоканальная Информационная Скоростная Система) с автономными контроллерами, вычитывающими регистрирующую электронику в режиме последовательного чтения информации и буферизующими данные во время сброса ускорителя. По концу сброса данные, накопленные в буферной памяти, вычитываются и обрабатываются. Пакет DATE формирует подсобытия из фрагментов событий и целые события из подсобытий посредством сети сборки событий. Сеть сборки событий с топологией звезда построена на основе 1GbE интерфейсов и 24x1GbE коммутатора. В основе распределённой системы хранения лежит кластерная файловая система GlusterFS.

**Abstract**

Donskov S.V. et al. Data Acquisition System for the Experiment OKA: IHEP Preprint 2013-22. – Protvino, 2013. – p. 28, figs. 8, tables 3, refs.: 34.

Data acquisition system of experiment OKA dedicated to research of rare kaon decays on accelerator U-70 is described in the paper. MISS system with autonomous controllers to read out front-end electronics by sequential data read and buffer data for accelerator burst serves as the basis of a blocked DAQ. Data buffered in controller memory are transfered and processed at the end of accelerator burst. DATE package builds subevents from event fragments and events from subevents with event building network. Event building network with topology star is made with 1GbE interfaces and 24x1GbE switch. Cluster file system GlusterFS is used to build a distributed storage system.

# Введение

Изучение редких распадов каонов на установке ИСТРА (У-70, Протвино) показало перспективность выбранного направления исследований. По результатам обработки данных, накопленных за сеансы в 2001-2003 годах, впервые в мире было получено заметное улучшение в точности параметров распадов каонов, что было опубликовано в работах [1], [2], [3], [4], [5]. Результатом публикаций стало решение провести эксперимент с усовершенствованным каналом частиц и улучшенными параметрами установки. Для сепаратора каонного пучка, использованного на $\Omega-$спектрометре в ЦЕРН и перевезённого в ИФВЭ, была разработана и построена уникальная криогенная система, что позволило иметь при интенсивности $3 \cdot 10^6$ частиц в пучке до $8 \cdot 10^5$ каонов в канале установки во время сброса ускорителя [6]. При распаде 15% каонов в распадной базе установки возможно иметь темп приёма триггерных событий в системе сбора данных до 120 кГц. Планировавшееся увеличение интенсивности триггеров привело к необходимости разработки усовершенствованной регистрирующей электроники, а увеличение суммарного количества каналов электроники привело к необходимости использования распределённой системы сбора данных и хранения.

Структурно система сбора состоит из следующих уровней (по пути передачи данных):

- регистрирующая электроника, преобразующая сигналы детекторов;
- электроника, вычитывающая данные по триггеру и буферизующая их во время сброса;
- компьютеры, вычитывающие данные из буферной памяти электроники и программно обрабатывающие их с целью сборки подсобытий из фрагментов;
- компьютеры, собирающие целые события;

- архивация данных на распределённую систему хранения и ленточную библиотеку.

Первые два уровня рассматриваются в разделах “Регистрирующая электроника” и “Вычитывающая электроника”. Следующие два уровня описаны в разделе “Архитектура”. Последнему уровню посвящён раздел “Система хранения”. После описания системы сбора данных приведены измерения и анализ некоторых её характеристик. По результатам анализа сделаны выводы для устранения проблем.

## Регистрирующая электроника

Вся регистрирующая электроника выполнена на модулях в системе МИСС. Амплитудный анализ калориметрических данных установки реализован на модулях ЛЭ-71. Трековая часть установки регистрируется с помощью модулей ЛЭ-78, ЛЭ-84, ЛЭ-95, ЛЭ-76.

Время оцифровки зависит от типа модуля и может зависеть от количества преобразованных им сигналов (таблица 1).

| Наименование | Тип | Число каналов | Время оцифровки (мкс) | количество слов |
|---|---|---|---|---|
| ЛЭ-71 | ЗЦП | 96 | 5 | $N_{hits}$ |
| ЛЭ-78 | ВЦП | 64 | 6 | $3 + N_{hits}$ |
| ЛЭ-84 | ВЦП | 64 | 1 | $2 \cdot (2 + N_{hits})$ |
| ЛЭ-95 | ВЦП | 64 | 3.3 | $1 + N_{hits}$ |
| ЛЭ-69 | пересчётка | 16 | 0 | 34 |
| ЛЭ-76 | регистр | 64 | 0 | 4 |
| ЛЭ-79 | триггерный процессор | - | 0 | 3 |

Таблица 1. Характеристики регистрирующих модулей МИСС, существенные для системы сбора данных [7], [8], [9], [10], [11]. $N_{hits}$ – количество хитов в модуле, константа в количестве слов определяется форматными словами.

ЗЦП ЛЭ-71 содержат 96 интегрирующих каналов. Время интегрирования сигнала задаётся длительностью триггерного строба в диапазоне 20-200 нс, разрядность оцифровки данных составляет 12 бит, длительность преобразования составляет $\sim 5$ мкс. При обработке данных модулем пьедесталы вычитаются с помощью программируемой пьедестальной памяти индивидуально для каждого канала. 40 модулей ЛЭ-71 размещены в 4 каркасах МИСС и обеспечивают оцифровку 3840 каналов калориметрии.

ВЦП ЛЭ-78, реализованные на основе кольцевого буфера емкостью 512 ячеек с остановом записи по переднему фронту триггерного строба, имеют временное разрешение 5 нс. В состав установки входит 5 каркасов ВЦП ЛЭ-78 регистрирующих в сумме 6400 каналов. ВЦП ЛЭ-84 оцифровывает сигналы с помощью двух микросхем HPTDC, регистрирующих по 32 канала и запрограммированных на разрешение 0,2 нс. В системе сбора используется один каркас ВЦП ЛЭ-84, содержащий 1280 каналов. Один каркас ВЦП ЛЭ-95, состоящих из двух субмодулей по 32 канала и имеющих временное разрешение 2 нс, также регистрирует 1280 каналов.

Каркас с триггерным процессором содержит шесть регистровых модулей ЛЭ-76 (384 канала), два решающих модуля ЛЭ-79, пересчётку ЛЭ-69 (16 каналов) для монитора счётчиков и триггерных решений и регистр признаков триггера ЛЭ-76 (64 канала).

Модули в каркасе с триггерным процессором имеют времена оцифровки существенно меньшие микросекунды, поэтому мёртвое время в этом каркасе зависит только от количества передаваемых данных, которое постоянно.

## Вычитывающая электроника

Необходимость уменьшения мёртвого времени системы сбора данных привела к разработке автономных контроллеров магистрали МИСС с буферной памятью на цикл ускорителя.

После каждого триггера во время сброса ускорителя модули регистрирующей электроники передают данные по магистрали МИСС в буферную память автономного контроллера в режиме последовательного чтения информации (ПЧИ). После прихода триггерного сигнала контроллер ожидает появления сигнала готовности на магистрали, по которому запускает процедуру последовательного чтения информации из модулей. ПЧИ также запускается, если сигнал готовности не появляется в течение 10 мкс. Первоначальный вариант контроллера ЛЭ-85 содержит 32 МБ буферной памяти и вычитывает модули по асинхронному протоколу. Время передачи одного 28 битного слова (12 бит адреса и 16 бит данных) составляет 200 нс, что позволяет передавать данные со скоростью 17 МБ/с [12]. При переходе на синхронный протокол ПЧИ время передачи уменьшается до 100 нс, что позволяет передавать данные по магистрали со скоростью 35 МБ/с. С момента прихода триггерного сигнала и до завершения ПЧИ контроллер выдаёт сигнал “Busy” на разъём передней панели. Мёртвое время в каркасах МИСС является суммой времени преобразования (таблица 1) и времени передачи данных. Время передачи данных линейно зависит от количества слов и равняется в микросекундах: $t_{read} = 0.1 \cdot N_{words}$, где $N_{words}$ – количество передаваемых слов.

По завершению цикла ускорителя содержимое буферной памяти передаётся по 32-разрядной магистрали через адаптер ЛЭ-75 в интерфейсную карту PCI7200, устанавливаемую в компьютер [13]. PCI карта PCI7200 производителя ADLINK позволяет принимать данные по 32-разрядной магистрали с уровнями TTL [15]. При иници-

ализации драйвера система выделяет драйверу область системной памяти (буфер чтения) в единоличное пользование. Из встроенной аппаратной FIFO объёмом 32 байта принятые данные передаются в буфер чтения по шине PCI в режиме DMA. Драйвер от производителя позволяет запускать процесс передачи данных из FIFO в буфер чтения, контролировать количество переданных байт и осуществлять копирование данных из буфера драйвера в буфер пользовательской программы.

В каждый их трёх компьютеров вычитывания было установлено две интерфейсных карты. Каждая карта принимала данные от двух каркасов МИСС, включённых в ветвь. Средняя измеренная скорость приёма данных одной картой по асинхронному протоколу составляет 6 МБ/с.

Использование PCI7200 налагает ограничения:

- необходимость использования двоичных драйверов интерфейсной карты, предоставляемых производителем для ограниченного числа версий Linux;
- невозможность получения сигнала окончания передачи данных при асинхронном чтении через интерфейсные вызовы;
- недостаточная скорость передачи в 6 МБ/с на один интерфейс;
- ограничение на максимальный объём одной передачи в 64 МБ через один интерфейс [16];
- переход производителей материнских плат компьютеров с шины PCI на PCI Express (позднее был выпущен PCI Express интерфейс).

С целью устранения описанных выше ограничений в ОЭА был разработан автономный контроллер ЛЭ-97 с интерфейсом USB2.0 для передачи данных. ЛЭ-97 оборудован 64 МБ буферной памяти и допускает установку мезонинной платы для увеличения размера памяти до 128 МБ [14]. Интерфейсная часть для передачи данных в компьютер также реализована в виде мезонинной платы и допускает замену соединения по протоколу USB2.0 на соединение по другому протоколу без необходимости полной переделки контроллера.

Физический, канальный и протокольный уровни шины USB для устройства реализуются интерфейсной платой с помощью микросхемы CY7C68001, имеющей slave FIFO интерфейс для внешнего процессора [17]. Данная микросхема сконфигурирована для пересылки данных через один канал передачи типа “bulk in” по протоколу USB2.0 с буфером передачи 512 байт в режиме “high speed”. Выбранная конфигурация предназначена для пересылки больших объёмов данных и гарантирует их целостность.

Для приёма данных на стороне компьютера был разработан драйвер missusb для операционной системы Linux, позволяющий выполнять синхронное и асинхронное чтение USB устройства, переконфигурирование буфера приёма после загрузки драйвера и получение идентификатора устройства (драйвер реализует USB Device Layer

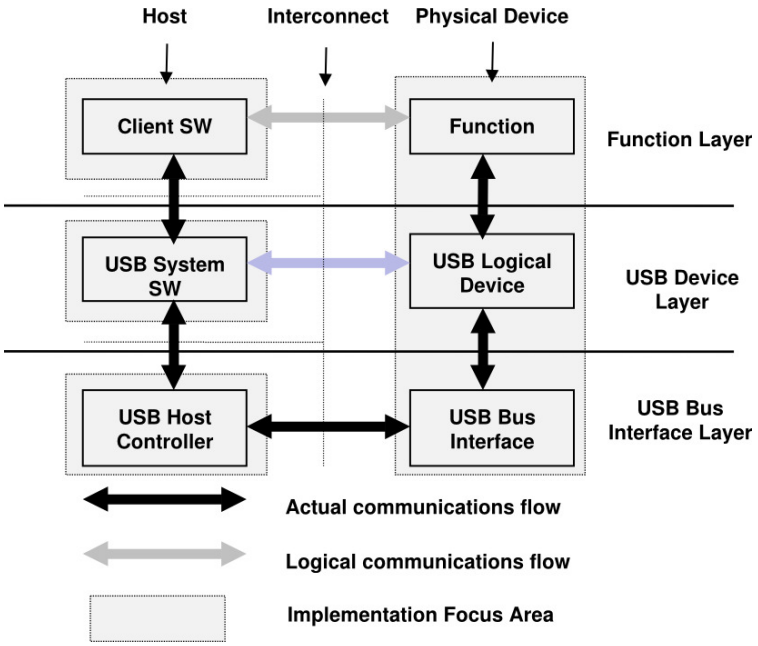


Рис. 1. Уровни реализации USB соединения [18].

на рис. 1). При синхронном чтении данные вычитываются драйвером в буфер передачи, затем копируются в буфер пользователя и цикл чтения в драйвере повторяется. При асинхронном чтении драйвер вычитывает данные в элементы буферной FIFO, откуда они копируются по вызову пользовательской программы в её буфер асинхронно с вычитыванием устройства. Размер буфера передачи, элементов FIFO и её длины конфигурируются пользовательской программой перед началом чтения.

Драйвер разрабатывался под семейство версий Linux 2.6.x и позднее был портирован для версий 3.x.y. Передача данных с шины USB в системную память происходит с использованием DMA. Количество одновременно подключаемых устройств к одному контроллеру USB2.0 ограничено протоколом USB и составляет 127 устройств. Уникальный адрес на шине USB устройство получает динамически при инициализации каждый раз заново, поэтому идентификация автономного контроллера по USB адресу невозможна. С целью возможности переконфигурирования системы сбора данных в конфигурацию устройству USB прошивается уникальный номер (DeviceID в default descriptor), используемый драйвером для наименования символьного файла устройства. Измеренная средняя скорость передачи данных из буфера контроллера в буфер пользовательской программы составляет 18 МБ/с и ограничена на стороне автономного контроллера.

Многоуровневая архитектура современных шин требует использования логических анализаторов шин при отладке устройств (USB имеет три уровня реализации: физический, канальный и протокольный). В противном случае необходимо изобретать изощрённые тесты с последующим анализом результатов их проведения и вы-

явлением причин проблем по косвенным признакам. Так причина одной из проблем в алгоритме работы разработанного контроллера, вызванная отсутствием пакета об окончании передачи при размере передаваемых данных кратному размеру буфера передачи, была выявлена и устранена через год опытной эксплуатации после нескольких итераций модификация-тестирование.

## Архитектура

Программной основой распределённой системы сбора данных послужил пакет DATE, разрабатываемый коллаборацией эксперимента ALICE (ЦЕРН). Данный пакет предполагает установку:

- компьютеров, вычитывающих электронику (Local Data Concentrator), осуществляющих предварительную сборку подсобытий из фрагментов и их передачу по протоколу TCP/IP сборщикам событий;
- компьютеров, принимающих подсобытия и осуществляющих конечную сборку событий (Global Data Collector);
- несколько вспомогательных компьютеров для управления и мониторинга данных.

Пакет DATE позволяет вычитывать электронику со скоростью до 1 Тб/с, осуществлять сборку событий с суммарной скоростью более 7,5 ГБ/с и передавать данные в хранилище с суммарной скоростью до 4,5 ГБ/с и предназначен для использования под операционной системой Linux в 32-битной версии [19].

Текущая конфигурация системы сбора данных установки ОКА включает три компьютера, вычитывающих электронику (КВЭ), и компьютер сборки событий (КСС), передающий данные на четыре узла кластерной файловой системы.

При работе системы сбора данных автономные контроллеры вычитывают регистрирующую электронику по магистрали МИСС во время сброса ускорителя, буферизуют данные и по окончанию сброса передают их в КВЭ.

Для подключения USB интерфейсов автономных контроллеров к USB контроллерам компьютеров используются кабели длиной 3, 5 и 7,5 метров с двойным экранированием производителя HAMA. Пассивные кабели длиной 7,5 метров (без повторителей), превышающие максимальную длину 5 метров по стандарту USB2.0 [18], не вызывают проблем при инициализации устройства на шине (bus enumeration) и демонстрируют стабильную передачу данных.

Один из компьютеров, вычитывающих электронику, получает прерывания от сигналов начала/конца сброса ускорителя и оповещает о них другие КВЭ с целью запуска процедуры вычитывания буферов автономных контроллеров. Вычитанные буферы програмно обрабатываются, из них извлекаются фрагменты событий. Заголовки фрагментов событий, записываемые автономными контроллерами, проверяются на

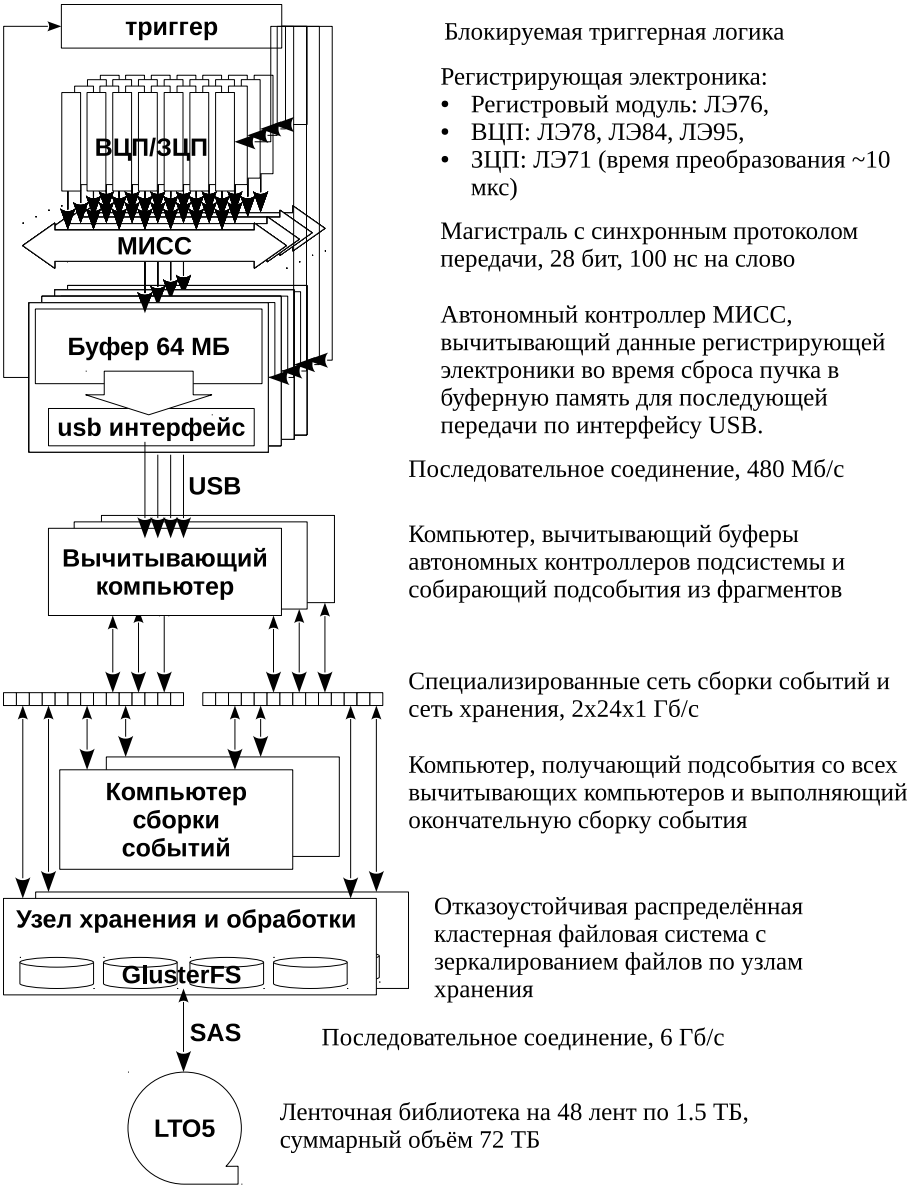


Рис. 2. Архитектура системы сбора данных установки ОКА.

соответствие формату и порядковому номеру события в сбросе. Суммарные количества фрагментов, переданных контроллерами, сравниваются между собой. При обнаружении сбоя при любой из проверок в заголовке подсобытия поднимается флаг, соответствующий типу сбоя. Из фрагментов собираются подсобытия, которые передаются КСС через сеть сборки событий (Event Building Network).

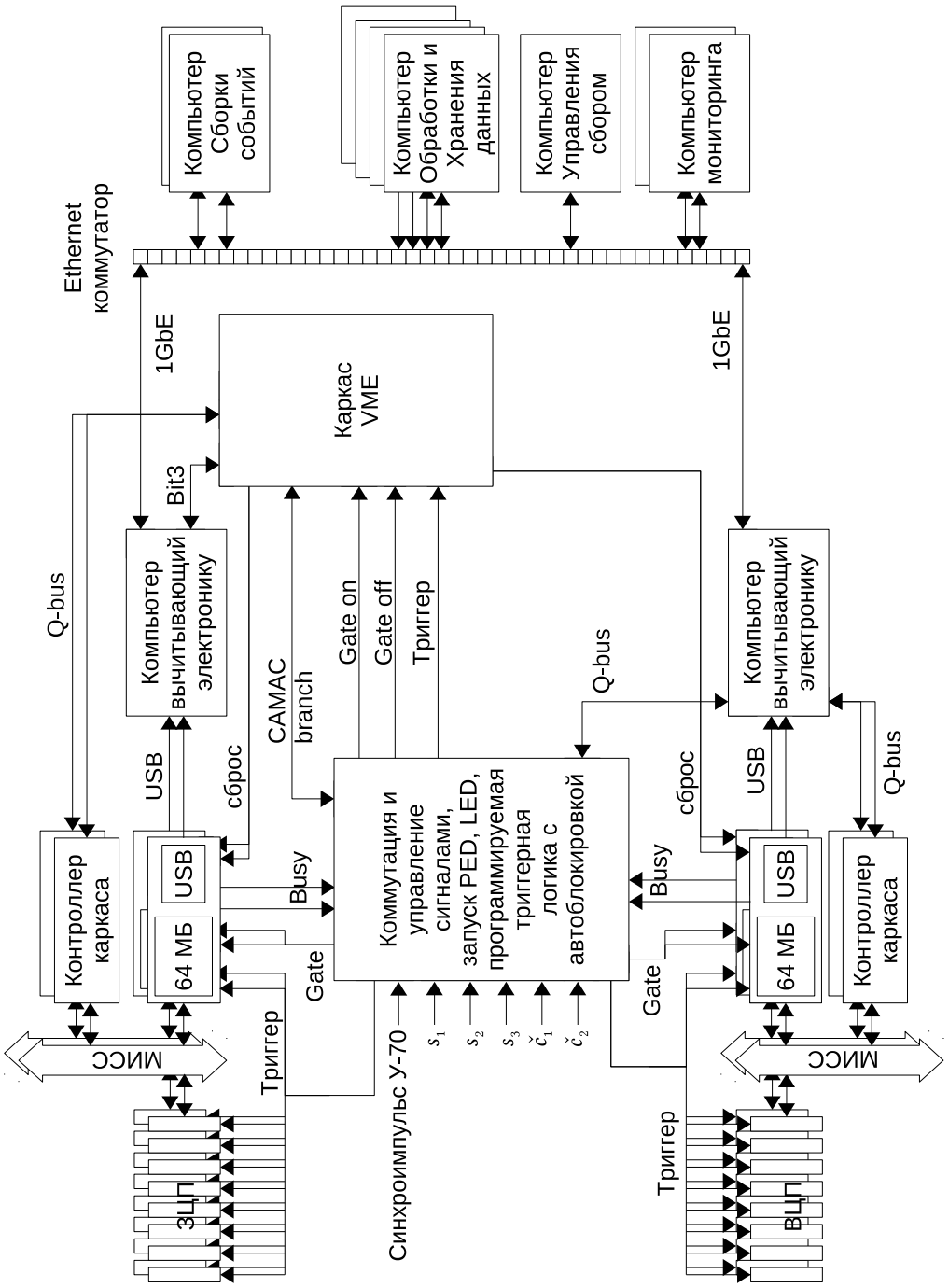


Рис. 3. Упрощённая структурная схема системы сбора данных установки ОКА.

Компьютеры системы сбора данных оборудованы двумя гигабитными интерфейсами каждый и объединены в сеть сборки событий с топологией звезда и такую

же независимую сеть хранения данных. Обе сети реализованы на неуправляемых коммутаторах D-Link DGS-1024D. Данный коммутатор имеет 24 гигабитных порта, обслуживаемых коммутационной фабрикой с пропускной способностью 48 Гбит/с [29].

КСС получает подсобытия с КВЭ, собирает события в соответствии с порядковым номером в сбросе и сохраняет их на локальный диск, откуда файлы с данными архивируются в систему хранения. Несколько компьютеров установки используются для управления системой сбора данных и мониторирования принимаемых данных установки.

Перед началом сброса ускорителя генерируются и вычитываются пьедестальные события. Вычисленные значения пьедесталов для каждого канала записываются в пьедестальные памяти и передаются системе сбора данных как пьедестальное событие. По окончанию сброса генерируется и вычитывается событие LED. Информация от пересчёток триггерной логики, интенсиметров, состояние программируемых блоков триггерной логики передаётся программой управления триггером в систему сбора данных посредством сегмента разделяемой памяти и добавляется в событие LED.

Для доступа к магистрали МИСС используются контроллеры ЛЭ-83, работающие в качестве ведомых устройств на кабельной шине Q-bus [20]. ЛЭ-83 используются для вычитывания пьедесталов ЛЭ-71, загрузки их пьедестальных памятей, тестирования и наладки модулей МИСС.

Интерфейс к шине Q-bus реализован на адаптерах шины V-07, разработанных в ОЭА и выполненных как VME модули. Доступ к шине VME на двух компьютерах организован с помощью адаптера Bit3, на третьем с помощью адаптера PCI-Qbus, разработанного в ОЭА [20]. Модуль Corbo, установленный в VME каркасе, позволяет принимать сигналы начала/конца сброса ускорителя через прерывания и подсчитывать количество триггеров в сбросе. Для доступа к каркасам CAMAC, в которых установлены модуль коммутации физических, PED и LED триггеров и генератор LED, используется контроллер ветви CAMAC CBD8210, выполненный как модуль VME.

После перехода на вычитывание контроллеров МИСС через USB интерфейсы компьютеры системы сбора данных были переустановлены под Linux 2.6.35 в 64-разрядной версии с целью снятия ограничения на объём доступной памяти и возможности полноценного использования современной вычислительной техники. Пакет DATE и драйверы missusb, Bit3 и PCI-Qbus были портированы на 64-битную систему.

## Система хранения данных

Планируемый объём принимаемых данных, исчисляемый десятками терабайт, потребовал выбора приемлемого варианта системы хранения. Система хранения должна была обеспечивать:

- требуемую производительность ввода-вывода;
- POSIX интерфейс, единое пространство имён директорий и файлов;
- возможность продолжения работы после отказа устройства хранения (обработка отказа жёсткого диска на лету);
- нетребовательность к аппаратным ресурсам, возможность использования обычных компьютеров, простоту в установке и поддержке.

Первоначально в качестве хранилища данных на установке OKA была выбрана дисковая стойка Promise UltraTrak SX8000 с аппаратным RAID контроллером и интерфейсом SCSI, подключенная к компьютеру со SCSI адаптером и NFS сервером. NFS сервер позволял монтировать разделы дисковой стойки на пользовательских компьтерах посредством локальной гигабитной сети. В ходе эксплуатации хранилища были изучены его характеристики, сильные и слабые стороны и к списку требований к системе хранения были добавлены пункты:

- масштабируемость системы хранения (линейность зависимости между суммарным объёмом хранилища и производительностью ввода-вывода);
- возможность увеличения объёма системы хранения без копирования хранимых данных во вспомогательное хранилище;
- возможность быстрого восстановления целостности данных после отказа устройства хранения не обработанного на лету.

В качестве альтернативных систем хранения были рассмотрены варианты на основе кластерных файловых систем. Архитектурно кластерные ФС состоят из узлов хранения, разделяемых по сети пользовательскими узлами. Сравнение систем хранения на основе аппаратных RAID контроллеров и сетевых ФС (“классические хранилища”) с кластерным ФС на обычных потребительских компьютерах (“кластерные хранилища”) показало, что:

- классические хранилища проигрывают кластерным по отношению стоимость-производительность. Кластерное хранилище может предполагать покупку только дисков для установки на кластерных узлах;
- классические хранилища масштабируются хуже кластерных, поскольку сетевые ФС обычно не рассчитаны на распределённую архитектуру;
- распределённый програмный RAID при многократном дублировании ненадёжных частей может быть не менее надёжен, чем аппаратный RAID;
- восстановление ФС после отказа устройства хранения не обработанного на лету в случае классического хранилища требует проверки всего раздела ФС. Кластерное хранилище может потребовать восстановления ФС только на узле с отказавшим диском;

- к явным преимуществам аппаратных RAID можно отнести простоту в установке и поддержке, а также большое время наработки на отказ.

Кластерные ФС активно развиваются в настоящее время и для обоснованного выбора требуется тщательный анализ их характеристик и условий их применения. Доступные варианты кластерных ФС возможно сравнить между собой по набору характеристик:

- полнота реализации интерфейсных вызовов POSIX;
- хранение метаданных и механизм управления блокировками;
- возможность избыточного хранения данных и обработки отказов на лету;
- масштабируемость ФС;
- расширяемость во время работы;
- лицензионная политика, поддержка со стороны разработчиков.

Одно из ключевых различий между кластерными ФС заключается в протоколе обмена по сети, его тип позволяет разделить кластерные ФС на два класса:

- ФС на основе сетевых блочных устройств (drbd, nbd, iSCSI, ATA over Ethernet и т.д.). ФС данного типа используют низкоуровневый протокол блочных устройств для обмена между узлами. Сеть в таком случае называют специализированной сетью хранения (Storage Area Network). Примеры: GFS, GPFS, OCFS;
- ФС на основе широко распространённых локальных ФС. ФС данного типа используют протокол уровня файлов для обмена между узлами. Примеры: Lustre, GlusterFS, PVFS2, Ceph [21]. К данному типу относятся распределённые хранилища на основе устройств хранения с сетевым интерфейсом (Network Attached Storage).

Кластерные ФС на основе локальных ФС позволяют иметь доступ к данным на узлах хранения напрямую через локальную ФС. Данная возможность позволила через несколько лет эксплуатации хранилища на основе такой кластерной ФС на установке OKA произвести идентификацию аппаратуры, работающей со скрытыми сбоями, нарушающими целостность данных (silent data corruption), и произвести полное восстановление данных без архивной копии. “Тихой порче данных” как явлению посвящено несколько исследований. Например, в ЦЕРН в работе [22] была дана оценка вероятности порчи данных на уровне $10^{-9}$. Серьёзность проблемы “тихой порчи” при хранении больших объёмов данных подтверждает факт разработки фирмой Sun Microsystems специальной ФС под названием ZFS, одной из главных целей создания которой являлась непрерывная проверка данных на целостность [23].

После изучения доступных кластерных файловых систем в качестве хранилища данных была выбрана GlusterFS с гигабитными интерфейсами на узлах хранения и пользовательских узлах [24]. Данная ФС способна масштабироваться до нескольких петабайт данных и обслуживать несколько тысяч клиентов одновременно. Для восстановления после отказов (split-brain problem) GlusterFS использует информацию, хранимую в расширенных файловых атрибутах (extended file attributes), поэтому в качестве локальной ФС, поддерживающей их хранение, была выбрана XFS, рекомендованная разработчиками. Для решения этой проблемы другие ФС, например GFS, используют сложные распределённые механизмы голосования и блокировок (quorum and fencing) [25].

GlusterFS поддерживает стандарт POSIX и доступна под лицензией GPL в исходных текстах. Архитектурно указанная ФС состоит из узлов хранения (storage bricks) и пользовательских узлов. Узел хранения содержит диски с разделами, отформатированными локальной ФС, которые посредством серверного процесса экспортируются пользовательским узлам. Серверный процесс на стороне клиента импортирует разделы узлов хранения и позволяет примонтировать их как единый объединённый раздел ФС посредством модуля операционной системы FUSE, обеспечивающего монтирование разделов ФС, реализуемых программами пользователя. Есть возможность импортировать разделы напрямую с помощью библиотеки libglusterfs для повышения производительности. Данная библиотека подменяет системные вызовы ввода-вывода при доступе к разделу GlusterFS с целью прямых обращений к узлам хранения, минуя FUSE.

В качестве транспорта для служебного протокола внутреннего обмена GlusterFS предоставляются на выбор TCP/IP, ориентированный на соединение, и OpenFabrics verbs RDMA-совместимые протоколы, такие как Infiniband, ориентированные на обмен сообщениями и реализующие удалённый прямой доступ к памяти с низкими задержками в обход ядра операционной системы (kernel bypass) [26]. Таким образом, в зависимости от решаемой задачи возможно использовать дешёвые гигабитные интерфейсы (1 Гбит/с, задержка $\approx$ 100 мкс), более дорогие и производительные десятигигабитные интерфейсы (10 Гбит/с, задержка $\approx$ 20 мкс) или ещё более дорогие и скоростные Infiniband контроллеры (10-40 Гбит/с, задержка $\approx$ 1 мкс) [27].

Сервисы узлов хранения и пользовательских узлов имеют гибкую архитектуру, перенастраиваемую с помощью конфигурационных файлов при загрузке сервиса. Базовые и дополнительные функции сервисов реализуются при помощи подгружаемых стекируемых модулей (translators), работающих в пространстве пользователя. Дублирование файлов с данными реализуется с помощью такого модуля и позволяет продолжать использовать ФС без ограничений в случае отказа одного узла хранения в дублирующей паре.

При эксплуатации системы хранения произошло несколько инцидентов вследствие аппаратных отказов. Например, во время одного из сеансов произошла поломка двух дисков в разных дублирующих парах (сервисы на узлах хранения номер 3 и 5 аварийно завершили выполнение из-за отказа дисков), первая пара дисков была

переполнена, и в таких условиях продолжалась нестабильная запись (рис. 4.А). После исключения из конфигурации дисков, отказавших во второй и четвёртой парах, и запуска сервисов стабильная запись была продолжена без ущерба для производительности (рис. 4.Б). Переконфигурирование кластерной ФС после аппаратных отказов не происходит автоматически на лету и требует ручного вмешательства в отличие от восстановления, которое требует только ручного запуска.

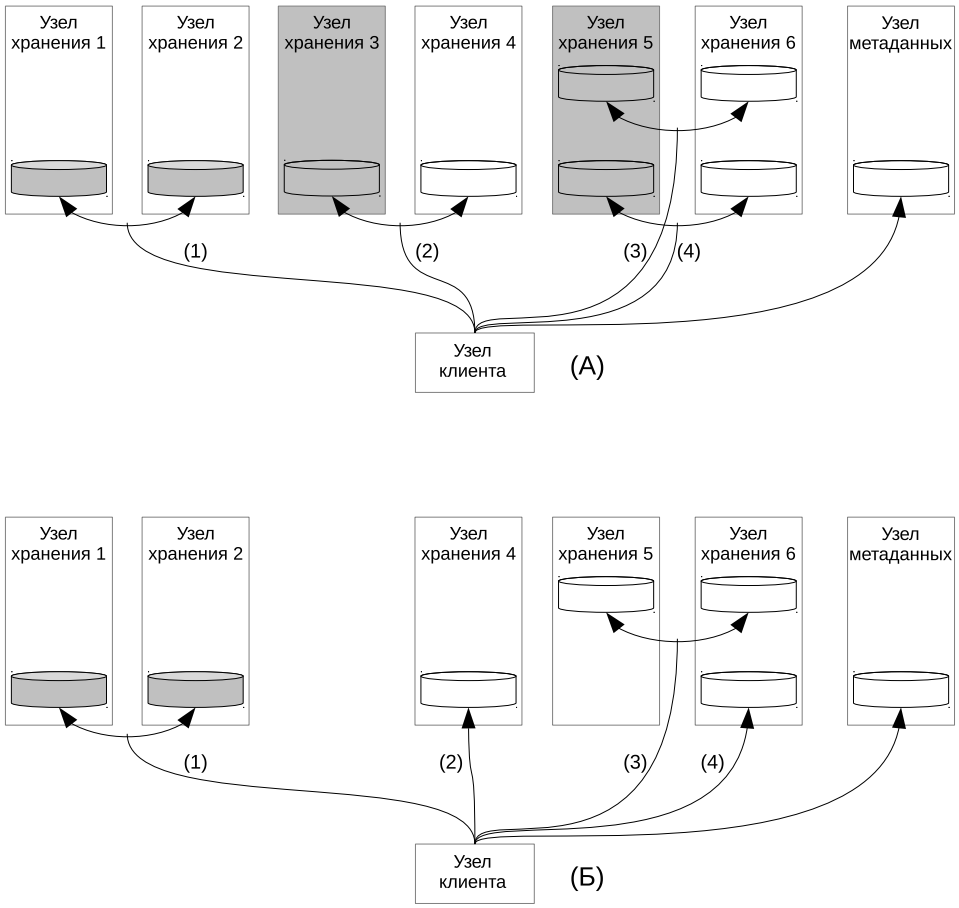


Рис. 4. Конфигурация GlusterFS в осеннем сеансе 2009 года после сбоев.

Во время пробной записи принимаемых данных на GlusterFS напрямую была обнаружена крайне низкая скорость передачи около 200 кБ/с. Было установлено, что причиной является запись данных пособытийно. После добавления буфера событий на 10 МБ скорость выросла до 80 МБ/с (два потока на две копии файла по 40 МБ/с). Вследствие большого числа отказов сетевых интерфейсов прямая запись была нестабильна, поэтому принимаемые данные вначале записываются на локальный диск. Разработанный сервис автоматически архивирует файл на GlusterFS и удаляет

его с локального диска. Архивация позволила уменьшить объём файлов с сырыми данными в 2,5 раза. С целью прямого чтения архивированных файлов библиотека доступа к данным пакета DATE была модернизирована. Тип файла с данными распознаётся при его открытии. Для архивированного файла автоматически запускается программа архивации, передающая разархивированные данные в программу пользователя напрямую без создания временной копии.

GlusterFS активно используется в некоторых центрах хранения и обработки данных, постоянно развивается командой разработчиков при непосредственном участии пользователей на протяжении более 6 лет. При изучении GlusterFS-2 в тестовой конфигурации на компьютере под операционной системой SLC4 была обнаружена проблема неправильной работы модуля распределения файла по узлам хранения (striping) вследствие особенностей интерфейса FUSE на старых ядрах. После контакта с разработчиками и дополнительных тестов по их просьбе, проблема была оперативно устранена в выпущенном патче.

Задокументированного способа расширять GlusterFS-2, добавляя новые узлы хранения, не предусмотрено, поэтому одним из авторов данной работы был изобретён способ делать это имеющимися средствами. В GlusterFS-3 данная проблема решается штатными средствами самой ФС.

За пять лет использования GlusterFS-2 её объём вырос до 9 ТБ (18 ТБ сырая ёмкость используемых дисков), из которых использовалось почти 7 ТБ. Подавляющее большинство сбоев в работе GlusterFS произошло из-за аппаратных отказов. Опытным путём была установлена крайняя ненадёжность встраеваемых и дискретных гигабитных интерфейсов на основе Ethernet контроллеров производителей Marvell и Realtek, практически каждый второй интерфейс отказал. Не зафиксированы отказы встраеваемых и дискретных интерфейсов на основе контроллеров производителей Intel и Broadcom.

Архитектура GlusterFS-2 предполагает обязательное использование одного раздела метаданных, обеспечивающего единое пространство имён. Данный компонент является “единой точкой отказа” (SPoF) и в случае поломки делает невозможным использование GlusterFS, поэтому его необходимо устанавливать на высоконадёжной аппаратуре. Третья версия GlusterFS избавлена от этого недостатка, используя “elastic hashing algorithm” вместо централизованой или распределённой модели хранения метаданных, что позволило достичь практически линейной масштабируемости ФС [28].

Надёжность контроллеров SATA-дисков, встроенных в материнские платы, оказалась невысокой. На одном из узлов хранения была выявлена порча данных контроллером с вероятностью $\approx 10^{-10}$ на всех подключенных к нему дисках при полном отсутствии системных сообщений об ошибках. Проблему удалось выявить благодаря проверке контрольных сумм, выполняемой архиватором gzip.

В начале 2013 года была создана новая система хранения на основе дискретных аппаратных RAID контроллеров 3ware SAS 9750-4i, установленных в четырёх новых узлах хранения, и GlusterFS третьей версии без дублирования файлов по узлам.

Измеренная скорость чтения RAID контроллеров составила 420 МБ/с, что вызвало необходимость модернизации гигабитных каналов передачи данных. Суммарное дисковое пространство нового хранилища данных доступное пользователям составило 44 ТБ, из которых на момент написания работы использовалось 18 ТБ.

Летом 2013 года были закуплены и установлены коммутатор с десятигигабитными портами HP 5406zl и десятигигабитные адаптеры Intel X540-T1 на новых узлах хранения, что позволило эффективно использовать аппаратные RAID контроллеры. Рост объёма принимаемых данных потребовал увеличения вычислительных мощностей по обработке реальных данных и генерации данных Монте-Карло, поэтому увеличенная суммарная пропускная способность хранилища и его объём позволили начать расширение вычислительного кластера на установке и организацию прямого десятигигабитного канала к вычислительному кластеру и резервному хранилищу данных в вычислительном центре ИФВЭ.

Расширение системы сбора и хранения данных привели к необходимости перехода на оборудование в 19-дюймовом стоечном конструктивном исполнении с целью повышения плотности размещения аппаратуры и облегчения её обслуживания, что заставило начать модернизацию инфраструктуры, включая питание, охлаждение и размещение оборудования и каналов передачи данных.

## Измерение и анализ характеристик системы сбора

Скорость вычитывания каркаса МИСС в ПЧИ при использовании асинхронного протокола составляла 17 МБ/с, что давало суммарную скорость вычитывания двенадцати каркасов $12 \cdot 17 = 204$ МБ/с без учёта времени преобразования. Время вычитывания каркасов занимало большую часть мёртвого времени, что привело к разработке более быстрого синхронного протокола ПЧИ. При переходе на синхронный протокол скорость вычитывания достигла 35 МБ/с, что дало суммарную скорость вычитывания каркасов $12 \cdot 35 = 420$ МБ/с.

Скорость передачи данных от автономного контроллера в компьютер через интерфейс PCI7200 в предыдущих сеансах составляла 6 МБ/с, что давало суммарную скорость передачи в компьютеры $6 \cdot 6 = 36$ МБ/с и ограничивало суммарный объём принимаемых данных. Наличие только двоичных сборок драйвера PCI7200 у производителя под устаревшие ядра и наличие интерфейса только под шину PCI не давало возможности использовать современные компьютерные компоненты. В качестве пробного варианта была разработана замена адаптера ЛЭ-75, преобразующего параллельную шину ЛЭ-85 в шину PCI7200. Новый адаптер ЛЭ-94 преобразовывал параллельную шину ЛЭ-85 в шину USB. После замены адаптера ЛЭ-75 и PCI7200 на адаптер ЛЭ-94 с интерфейсом USB скорость передачи выросла до 18 МБ/с на интерфейс, что давало суммарную скорость передачи $7 \cdot 18 = 126$ МБ/с и позволило начать минимизацию мёртвого времени системы сбора данных перед осенним сеансом 2011 года.

### Физический сеанс осенью 2011 года

Синхронный протокол ПЧИ и увеличившаяся в 3 раза скорость передачи данных на интерфейс позволила увеличить темп приёма данных. Средний размер события в осеннем сеансе 2011 г. составлял 4 кБ (рис. 5), что при среднем темпе набора 14 кГц давало $\sim 60$ МБ за сброс суммарно.

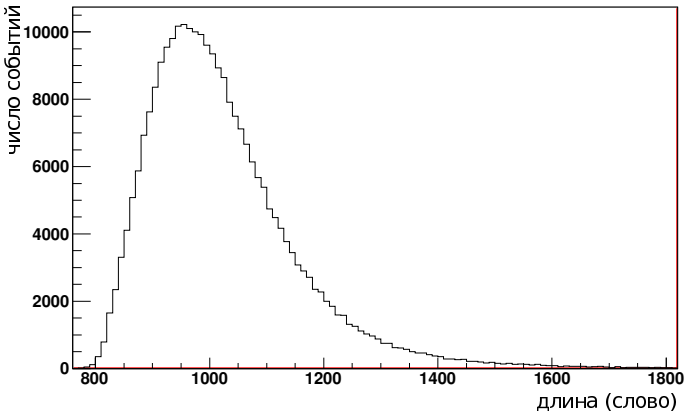


Рис. 5. Распределение размера события в словах в осеннем сеансе 2011 г.

В этом сеансе ЛЭ-95 имели 3 форматных слова на событие, каркас с триггерным процессором содержал две пересчётки ЛЭ-69. В таблице 2 приведены средние длины событий для каждого каркаса в осеннем сеансе 2011 года в наборе с физическим триггером.

Автономные контроллеры ЛЭ-85, объединённые в семь цепочек параллельными кабельными шинами, передавали данные в семь адаптеров ЛЭ-94, которые в свою очередь передавали данные в три компьютера через USB интерфейсы. Адаптеры ЛЭ-94 продемонстрировали безошибочную передачу данных по шине USB в течение сеанса, что привело к разработке автономного контроллера ЛЭ-97 с передачей данных в компьютер напрямую через интерфейс USB.

### Отладочный сеанс весной 2012 года

Перед коротким отладочным весенним сеансом 2012 года была проведена модернизация вычитывающей электроники. Двенадцать ЛЭ-85 и семь ЛЭ-94 были заменены на тринадцать ЛЭ-97, вычитываемых тремя компьютерами напрямую. Суммарная скорость передачи данных от автономных контроллеров в компьютеры выросла с $7 \cdot 18 = 126$ МБ/с до $13 \cdot 18 = 234$ МБ/с и окончательно перестала быть узким местом в системе сбора.

Контроллер ЛЭ-97 позволяет измерять мёртвое время каркаса электроники с точностью 100 нс в каждом событии. Это даёт возможность заниматься контролем и

| Номер каркаса | Наименование модулей в каркасе | длина события (слов) |
|---|---|---|
| 0 | ЛЭ-71 | 27,3 |
| 1 | ЛЭ-71 | 36,0 |
| 2 | ЛЭ-71 | 43,5 |
| 3 | ЛЭ-71 | 9,2 |
| 8 | ЛЭ-69, ЛЭ-76, ЛЭ-79 | 102,0 |
| 10 | ЛЭ-78 | 101,1 |
| 11 | ЛЭ-78 | 79,5 |
| 12 | ЛЭ-78 | 97,2 |
| 13 | ЛЭ-78 | 86,1 |
| 9 | ЛЭ-84 | 147,3 |
| 14 | ЛЭ-78 | 65,2 |
| 15 | ЛЭ-95 | 142,0 |

Таблица 2. Усреднённые характеристики каркасов МИСС в осеннем сеансе 2011 года.

настройкой мёртвого времени всей системы сбора и каждого каркаса по отдельности прямо во время приёма данных. С этой целью была добавлена таблица статусной информации каркасов МИСС в программу контроля принимаемых данных (рис.6). Программа контроля осуществляет раскодирование данных и выполняет функции:

- проверяет раскодированные данные на соответствие формату по описаниям модулей в каждом бите данных и отслеживает более 30 типов ошибок;
- накапливает статистику ошибок и длину данных для каждого модуля в системе сбора;
- отображает список модулей, отсортированный по частоте ошибок, для быстрого выявления наиболее проблемных модулей;
- выводит краткую информацию по количеству ошибок, длины данных, мёртвому времени для каждого каркаса с усреднением отдельно по набору и последнему сбросу;
- сортирует каркасы по мёртвому времени в каждом событии и накапливает процент событий, в которых каждый каркас занимал определённое место в отсортированном списке и процент ошибок на нём;
- содержит анализатор формата данных для побитового изучения слов события на предмет сбоев и смены формата данных в модуле, что позволяет быстро диагностировать проблемы в формате данных. При чтении потока данных можно задать критерий останова по типам ошибок, обнаруживаемых в каждом модуле, для поиска таких событий;

- при удовлетворении заданных критериев, таких как превышение пороговой средней длины события и порогового среднего мёртвого времени, выдаются звуковые оповещения операторам системы сбора для привлечения их внимания к возникшим проблемам.

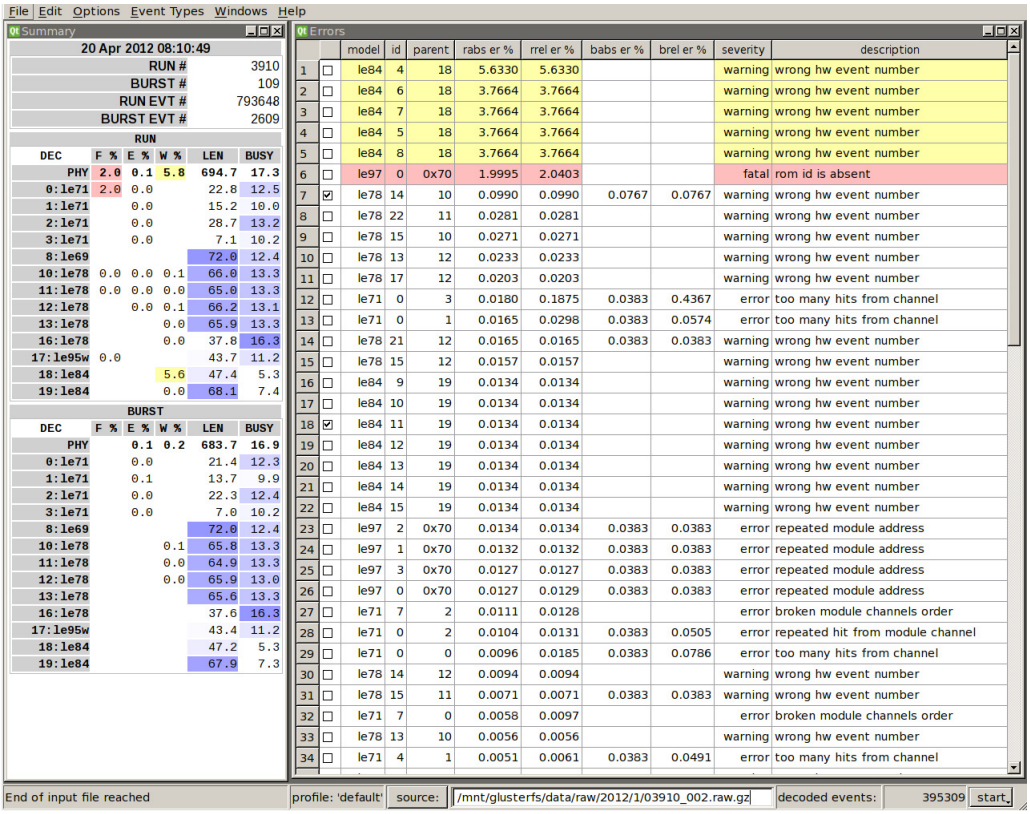


Рис. 6. Снимок экрана с интерфейсом программы контроля принимаемых данных.

Перед коротким отладочным весенним сеансом 2012 года была проведена работа по уменьшению мёртвого времени в каркасах с пересчётками ЛЭ-69, ВЦП ЛЭ-84 и ЛЭ-95, которые вносили основной вклад в суммарное мёртвое время системы сбора. Одна пересчётка, используемая для признаков триггера, была заменена на регистр ЛЭ-76, каркас с ЛЭ-84 был разделён на два независимых сектора, количество форматных слов в ЛЭ-95 было уменьшено с трёх до одного на событие, что существенно уменьшило мёртвые времена в этих карасах.

После модернизации в системе сбора основной вклад в суммарное мёртвое время в весеннем сеансе 2012 года стали вносить каркасы ЛЭ-78 и ЛЭ-71. Среднее мёртвое время максимальное среди каркасов на событие составляло 17,3 мкс, что при темпе набора 14 кГц давало бы суммарное мёртвое время каркасов около 24%.

| Номер каркаса | Наименование модулей в каркасе | длина события (слов) | мёртвое время (мкс) |
|---|---|---|---|
| 0 | ЛЭ-71 | 22,8 | 12,5 |
| 1 | ЛЭ-71 | 15,2 | 10,0 |
| 2 | ЛЭ-71 | 28,7 | 13,2 |
| 3 | ЛЭ-71 | 7,1 | 10,2 |
| 8 | ЛЭ-69, ЛЭ-76, ЛЭ-79 | 72,0 | 12,4 |
| 10 | ЛЭ-78 | 66,0 | 13,3 |
| 11 | ЛЭ-78 | 65,0 | 13,3 |
| 12 | ЛЭ-78 | 66,2 | 13,1 |
| 13 | ЛЭ-78 | 65,9 | 13,3 |
| 16 | ЛЭ-78 | 69,2 | 16,3 |
| 17 | ЛЭ-95 | 43,7 | 11,2 |
| 18 | ЛЭ-84 | 47,4 | 5,3 |
| 19 | ЛЭ-84 | 68,1 | 7,4 |

Таблица 3. Усреднённые характеристики каркасов МИСС в весеннем сеансе 2012 года.

### Анализ работы ПЧИ в весеннем сеансе 2012 года

Анализ мёртвых времён каркасов приводится отдельно для событий без сбоев и со сбоями во фрагментах событий, т.к. это позволило выявить некоторые особенности. Количество событий со сбоями в анализируемом наборе ($\sim 7\%$) не является типичным для набора физических данных ($\sim 0,5\%$). Анализ сопровождается предположениями о причинах обнаруженных особенностей.

Анализ корреляций длин событий с мёртвыми временами каркасов позволил изучить работу ПЧИ. В общем случае следует ожидать распределение мёртвых времён в виде отрезка, лежащего на наклонной прямой со следующими особенностями:

- наклон прямой соответствует скорости передачи во время ПЧИ;
- точка пересечения прямой с осью $y$ соответствует времени преобразования. Пересечение начала координат соответствует нулевому времени преобразования;
- начало отрезка соответствует количеству форматных слов в модуле. В случае выдачи форматных слов даже без наличия хитов начало отрезка имеет ненулевую координату по оси $x$.

ЛЭ-71 характеризуется отсутствием форматных слов и позволяет напрямую измерить время преобразования (рис. 7). Время преобразования данного модуля по спецификации составляет 5 мкс, тогда как измеренная задержка начала передач составляет 9-10 мкс. Данный факт объясняется процедурой вычитания пьедесталов, время которой не входит в указанное время преобразования и намеренно завышено в несколько раз для надёжной работы.

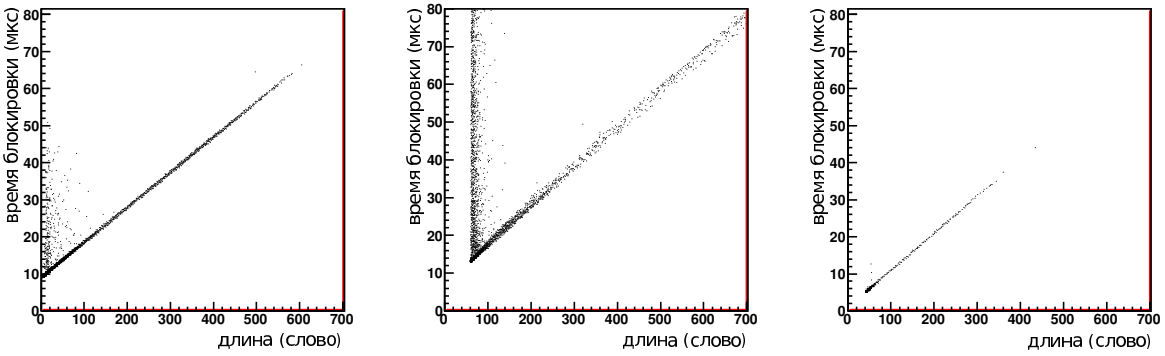


Рис. 7. Длина события на мёртвое время каркасов ЛЭ-71, ЛЭ-78, ЛЭ-84 в весеннем сеансе 2012 г.

Линейно интерполированная задержка перед вычитыванием для ЛЭ-84 составляет около 1 мкс. Точно определить её затруднительно, поскольку данный модуль даёт очень большое количество форматных слов. Интерполяция задержек начала передач для остальных модулей составляет: ЛЭ-78 около 7 мкс, ЛЭ-95 около 7 мкс, что согласуется с описаниями.

Скорость передачи по магистрали МИСС для модулей разных типов составляет 9-11 слов в микросекунду, что согласуется со значением 100 нс на слово, заявленным разработчиками.

Передача данных модулями ЛЭ-78 имеет следующую особенность. При небольшом числе хитов существует заметный разброс мёртвого времени (превышение в несколько раз от корректного значения) в 1,8% событий. Поскольку раскодирование проходит без ошибок, то количество слов, указываемое модулем в заголовке, корректное. Эту особенность можно объяснить внутренним устройством автономного контроллера:

- ЛЭ-97 содержит промежуточный буфер размером 1000 слов, предназначенный для организации триггера второго уровня. После вычитывания модулей в промежуточный буфер возможно обработать сигнал триггера второго уровня и по его наличию или отсутствию отбросить вычитанные данные или переместить их в основную память;

- ЛЭ-97 снимает сигнал busy до окончания пересылки данных из промежуточного буфера в основную память, что позволяет уменьшить мёртвое время каркаса.

Таким образом пересылка данных из промежуточного буфера в основную память может не успеть завершиться до прихода следующего триггерного сигнала, что приведёт к ожиданию завершения пересылки предыдущего события перед запуском ПЧИ для текущего.

Рассмотрим детально корреляцию длины события, раскодированного со сбоями, с его мёртвым временем в модулях ЛЭ-78. На этом графике появляется ромбовидный

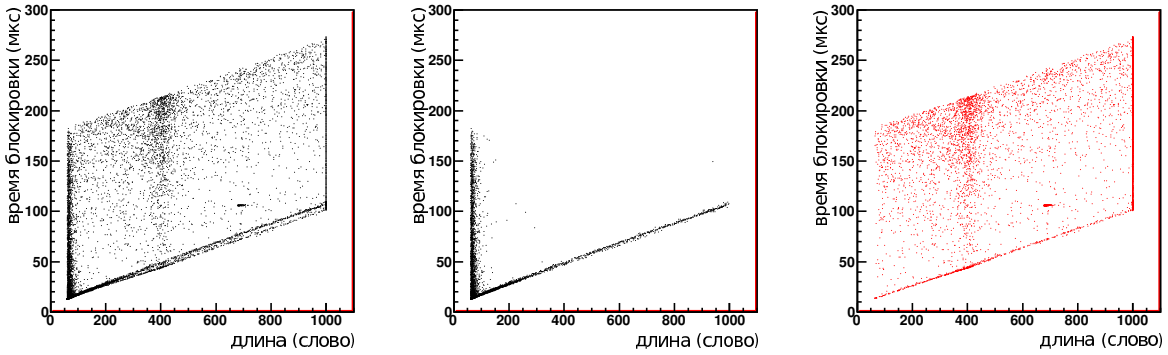


Рис. 8. Длина события на мёртвое время каркасов ЛЭ-78 в весеннем сеансе 2012 г для всех событий, без ошибок, с ошибками.

разброс мёртвого времени (рис. 8). Эту особенность можно объяснить внутренним устройством регистрирующих модулей:

- регистрирующие модули имеют внутреннюю память, которая в некоторых типах модулей не сбрасывается по приходу следующего триггерного сигнала (ЛЭ-78, ЛЭ-84, ЛЭ-95).

Таким образом при переполнении триггерного буфера данные, невычитанные в ПЧИ, могут остаться в памяти модулей и быть вычитанными в последующих ПЧИ, увеличивая их длительность, смешивая куски разных событий и ломая формат последующих событий. Время копирования предыдущего события в основную память даёт разброс по оси $y$, в соответствии с особенностями работы ЛЭ-97. Мёртвое время в таком случае может вырасти в несколько раз, что может привести к потере нескольких следующих триггерных событий. Данное явление актуально во время массового срабатывания каналов при самовозбуждении аналоговой электроники или наводках, что самоустраняется или устраняется персоналом через некоторое время после обнаружения, но регулярно происходит и приводит к заметному уменьшению темпа приёма триггерных событий (до двух раз и более). Представлялось необходимым реализовать быстрый сброс памяти модулей либо сигналом от контроллера по переполнению триггерного буфера, либо по приходу следующего триггерного сигнала.

Во время набора данных вычитывание контроллеров ЛЭ-97 по USB иногда останавливалось без видимых причин, что приводило к останову набора данных. Попытки выяснить причину оканчивались безрезультатно. Поскольку проблема проявлялась на контроллерах случайно и независимо, то её вероятность и соответственно важность возрастала с ростом числа контроллеров. Разработанные после сеанса дополнительные тесты позволили вскрыть причину явления. Было установлено, что останов передачи жёстко привязан к размеру передаваемых данных, который в случае останова был кратен размеру буфера передачи микросхемы CY7C68001, реализующей USB интерфейс в автономном контроллере. Возникло предположение об отсутствии последнего пакета с нулевой длиной данных на шине USB и, как следствие,

необходимости контроля заполненности буфера при посылке такого пакета. Модифицированная в соответствии с предположением прошивка автономного контроллера успешно прошла все тесты. Поскольку шина USB имеет многоуровневую архитектуру, то обычные инструменты разработчика, такие как осциллограф, не могут дать картины происходящего на верхних уровнях. Проблема с остановом передачи могла быть выявлена гораздо раньше и быстрее при наличии логического анализатора шины USB.

### Физические сеансы осенью 2012 года и весной 2013 года

Автономные контроллеры с модифицированной прошивкой вычитывались во время сеанса осенью 2012 года без проблем, что окончательно подтвердило устранение причины останова передач. Реализованный в прошивке автономного контроллера быстрый сброс модулей при возникновении ошибки во время ПЧИ позволил устранить её влияние на последующие события при массовом срабатывании каналов. Было набрано более 10 ТБ сырых данных, что стало рекордом для установки и для экспериментов на У-70 в целом. Основной вклад в мёртвое время системы сбора вносили каркасы ЛЭ-78, что привело к необходимости переработки их формата данных и уменьшения времени ПЧИ.

Перед сеансом весной 2013 года формат данных ЛЭ-78 был переработан, что позволило уменьшить количество форматных слов с трёх до одного на каждый модуль. Мёртвое время каркасов ЛЭ-78 уменьшилось с 13-16 мкс до 11-12 мкс и перестало определять мёртвое время системы сбора данных, которое стало зависеть от мёртвого времени каркасов ЛЭ-71. Средний темп приёма реальных данных составил 20 кГц при 50% суммарном мёртвом времени, а при запуске от генератора 54 кГц при 100% суммарном мёртвом времени. Коэффициент сжатия сырых данных упал с 2,5 до 2. В процессе тестирования модифицированных ЛЭ-78 была обнаружена и нивелирована некорректная работа одной из стандартных библиотечных функций разработчика ПЛИС Altera.

## Перспективы развития системы сбора данных

Дальнейшее развитие системы сбора данных вызвано необходимостью максимально эффективного использования возможностей имеющегося каонного пучка, и решением проблем возникающих при физическом анализе данных.

### Краткосрочный план развития

Исходя из цикла ускорителя длительностью 9 секунд и сброса длительностью 1 секунду на вычитывание буферов, их обработку и передачу остаётся 8 секунд. Максимальная суммарная измеренная скорость передачи по USB интерфейсам на компьютере с одним USB2.0 контроллером составляет 46 МБ/с. Скорость передачи

по сети сборки событий примем равной 80 МБ/с. Для надёжной работы требуется 1.5-2 кратное отношение теоретических пиковых значений к средним, то есть на вычитывание, обработку и пересылку буферов остаётся 4 секунды. Если принять, что время обработки буферизованных данных пренебрежимо мало по сравнению с временем ввода-вывода, то максимальный размер данных в буфере $V$ задаётся уравнением:

$$\frac{4 \cdot V}{46} + \frac{4 \cdot V}{80} = 4. \quad (1)$$

Откуда максимальный объём данных в буфере автономного контроллера получается равным 29 МБ = 7,3 мегаслова. При среднем размере события 100 слов, что с запасом соответствует реальным данным, получаем 73 тысячи событий за сброс. При увеличении длительности сброса до 2 секунд при темпе 50 кГц скорость передачи данных по USB интерфейсу и по гигабитному соединению может стать узким местом. Для увеличения пропускной способности данных соединений можно предпринять действия:

- установка дополнительного USB2.0 контроллера на каждом КВЭ позволит поднять суммарную скорость вычитывания автономных контроллеров до 90 МБ/с;
- при объединении нескольких сетевых интерфейсов в один виртуальный с помощью агрегирования соединений можно теоретически кратно увеличивать скорость передачи данных. Данная возможность обеспечивается Linux драйвером “bonding” [32]. Используемые коммутационные фабрики обеспечивают полную скорость передачи во всех направлениях одновременно, стандарт 1GbE обеспечивает полнодуплексное соединение [33], поэтому агрегированием соединений сети сборки событий и сети хранения можно обеспечить двукратный прирост скорости в обеих сетях одновременно при условии, что пересылки в этих сетях разнонаправленны в сетевых интерфейсах.

Это позволит принимать данные за двухсекундный сброс при темпе 50 кГц и вычитывать до $10^5$ событий за сброс. При этом суммарный объём буферов автономных контроллеров при среднем размере события 4 кБ составит 400 МБ за сброс и потребует увеличения объёма системы хранения. Объём архивированных данных за 2 недели непрерывного набора при степени сжатия 2 будет составлять 27 ТБ, а с учётом данных Монте-Карло 54 ТБ, что потребует увеличения объёма хранилища данных. Проводимая в настоящее время модернизация систем сбора данных и хранения обеспечит требуемые характеристики.

Канал обмена данными с контроллерами каркасов МИСС ЛЭ-83 посредством шины Q-bus, используемый для чтения пьедесталов, записи пьедестальных памятей в ЗЦП ЛЭ-71 и для отладки и тестов электроники, требует замены. Максимальная длина кабельной шины и количество устройств на ней уже вызывают проблемы при работе. Проводимая в настоящее время авторами данной работы разработка контроллеров каркасов МИСС с интерфейсом USB призваны решить описанную проблему.

### Проблемы при физическом анализе данных

Уменьшение порога триггерной сборки в ГАМС с 3-4 ГэВ до 1-2 ГэВ за счёт поднятия напряжения с 1,7 кВ до 1,8 кВ для уменьшения наводок привело к увеличению амплитуды в 2 раза и выходу сигнала за диапазон оцифровки от наиболее энергичных $\gamma-$квантов. Повышение разрядности оцифровки с 12 до 13 бит устранило бы означенную проблему.

### Долгосрочный план развития

Все используемые модули регистрирующей электроники за исключением ЛЭ-84 на основе HPTDC не позволяют кардинально уменьшить мёртвое время преобразования сигналов. Для достижения мёртвого времени 1 мкс с целью максимально эффективного использования пучка необходима их полная замена. Модернизированные модули в каркасах ЕвроМИСС могут обеспечить требуемые характеристики:

- модули ВЦП ЕМ-4 могут быть переделаны под конфигурируемое временное окно регистрации сигналов и конвеерную архитектуру без блокировки регистрации на время вычитывания модулей;
- модули ЗЦП ЕМ-6 имеют время преобразования 1 мкс.

Модернизация электроники с целью уменьшения мёртвого времени и, как следствие, возросший темп набора данных вместе с имеющимся резервом повышения интенсивности каонного пучка приведёт к увеличению объёма принимаемых данных в несколько раз, что потребует увеличения скорости вычитывания регистрирующей электроники, сборки и сохранения принятых данных.

Передачу большего объёма данных с регистрирующих модулей можно осуществить заменой магистрали МИСС (ЕвроМИСС) на последовательное соединение с каждым модулем. Разработчиками из ОЭА было предложено буферизовать данные в каждом модуле и вычитывать модули компьютером напрямую через USB. Замена соединения USB2.0 соединением USB3.0, позволит увеличить скорость с 18 МБ/с до 300 МБ/с на одно соединение [30]. Коммерчески доступные микросхемы CYUSB3014 реализуют физический, канальный и протокольный уровни шины USB3.0 со скоростью передачи 5 Гб/с и имеют slave FIFO интерфейс к внешнему процессору [31]. Объединение протоколов USB2.0 и USB3.0 на аппаратном уровне невозможно, поэтому интерфейс модулей должен сразу реализовываться на USB3.0.

Сборка и сохранение данных становятся всё более серьёзной проблемой при увеличении потока принимаемых данных. Зависимость стоимости требуемых вычислительных мощностей и каналов передачи данных от их пропускной способности растёт быстрее линейной при переходе к скоростям в сотни мегабайт и гигабайты принимаемых данных в секунду. Накладные расходы на обслуживание, питание, охлаждение таких систем сбора данных становятся отдельными проблемами, требующими серьёзного изучения. Одним из результатов такого изучения стала модернизация системы

сбора данных эксперимента с фиксированной мишенью COMPASS на ускорителе SPS в ЦЕРН. Существующая система сбора эксперимента COMPASS позволяет принимать данные со скоростью 2 ГБ/с. Вычислительные ресурсы этой системы сбора насчитывают около 30 компьютеров, вычитывающих электронику, и 20 компьютеров, собирающих события и локально буферизующих их посредством аппаратных RAID контроллеров. С целью минимизации парка вычислительной техники, накладных расходов и повышения надёжности распределённая програмная модель сборки событий, основанная на гигабитных каналах передачи, заменяется в настоящее время на распределённую аппаратную модель с програмным контролем посредством пакета IPBus. При этом количество компьютеров необходимых для вычитывания и локальной буферизации данных было уменьшено до 8 без уменьшения пропускной способности системы сбора данных [34].

С целью максимально эффективного использования вычислительных мощностей и каналов передачи данных одним из авторов настоящей работы разработана новая програмная модель сборки событий снижающая требования к вычислительной технике по сравнению с моделью, используемой в пакете DATE, и позволяющая оценочно достичь скоростей сборки событий в несколько ГБ/с на имеющихся вычислительных мощностях, что будет востребовано при уменьшении мёртвого времени системы сбора данных до микросекунды и реализации вычитывания электроники посредством протокола USB3.0.

## Заключение

В настоящей работе рассмотрена распределённая система сбора данных эксперимента по поиску редких распадов каонов OKA. Усовершенствованный канал частиц и, как следствие, высокая интенсивность триггерного сигнала потребовали разработки новой регистрирующей и вычитывающей электроники. Разработанный автономный контроллер МИСС с интерфейсом USB2.0 позволил в три раза увеличить скорость вычитывания буферной памяти по сравнению с промышленной картой цифрового ввода-вывода ADLink PCI7200 и устранить зависимость от двоичных драйверов производителя. Увеличившееся количество каналов детекторов и, как следствие, большой поток данных потребовали использования распределённых систем сбора данных и хранения. В качестве програмной основы для распределённой сборки событий был использован пакет DATE. Кластерная файловая система GlusterFS послужила основой для системы хранения данных.

В заключение авторы выражают благодарность В.Ф. Куршецову, В.В. Молчанову за помощь в проведении некоторых тестов и ценные советы по усовершенствованию системы сбора данных.

Работа выполнена при поддержке гранта РФФИ 11-02-00870-а.

## Список литературы


[1] O. P. Yushchenko, S. A. Akimenko, G. I. Britvich, K. V. Datsko, A. P. Filin, A. V. Inyakin, A. S. Konstantinov and V. F. Konstantinov *et al.*, "High statistic measurement of the K- —> pi0 e- nu decay form-factors," Phys. Lett. B **589**, 111 (2004) [hep-ex/0404030].

[2] O. P. Yushchenko, S. A. Akimenko, K. S. Belous, G. I. Britvich, I. G. Britvich, K. V. Datsko, A. P. Filin and A. V. Inyakin *et al.*, "High statistic study of the K- —> pi0 mu- nu decay," Phys. Lett. B **581**, 31 (2004) [hep-ex/0312004].

[3] I. V. Ajinenko, S. A. Akimenko, G. I. Britvich, K. V. Datsko, A. P. Filin, A. V. Inyakin, A. S. Konstantinov and V. F. Konstantinov *et al.*, "Measurement of the Dalitz plot slope parameters for K- —> pi0 pi0 pi- decay using ISTRA+ detector," Phys. Lett. B **567**, 159 (2003) [hep-ex/0205027].

[4] V. I. Romanovsky, S. A. Akimenko, G. I. Britvich, K. V. Datsko, A. P. Filin, A. V. Inyakin, A. S. Konstantinov and I. Y. Korolkov *et al.*, "Measurement of K- —> pi0 e- anti-nu(gamma) branching ratio," arXiv:0704.2052 [hep-ex].

[5] V. A. Duk *et al.* [ISTRA+ Collaboration], "Search for Heavy Neutrino in $K^- - > \mu^- \nu_h(\nu_h - > \nu\gamma)$ Decay at ISTRA+ Setup," Phys. Lett. B **710**, 307 (2012) [arXiv:1110.1610 [hep-ex]].

[6] "Эксперименты с заряженными каонами на сепарированном каонном пучке ускорителя ИФВЭ", предложение по прецизионному измерению каонных распадов на У-70 ИФВЭ, 2003.

[7] "QDC96", М. Солдатов, техническое описание

[8] "МОДУЛЬ 64-КАНАЛЬНОГО РЕГИСТРА ЛЭ-78", Васильев М.В., техническое описание.

[9] "Быстродействующие многоканальные модули ВЦП пикосекундного разрешения с программируемыми параметрами", Карпеков Ю.Д., Киселев Ю.С., Сенько В.А., Препринт ИФВЭ 2011-20, Протвино, 2011.

[10] "ЛЭ-95 – 64-канальный ВЦП в системе МИСС", Н. Шаланда, техническое описание..

[11] "Быстродействующий триггерный спецпроцессор для выделения распада частицы по координатной информации с годоскопов сцинтилляционных счётчиков", Н.С.Иванова и др., Препринт ИФВЭ 2007-12, Протвино, 2007.

[12] "Быстродействующая система регистрирующей и триггерной электроники для экспериментальных исследований в ИФВЭ", Бушнин Ю.Б., препринт ИФВЭ, 1988.

[13] “Автономный контроллер МИСС (ЛЭ-85) с возможностью буферизации информации за сброс ускорителя”, В. Якимчук, техническое описание.

[14] “Автономный контроллер МИСС (ЛЭ-97) с возможностью буферизации информации за сброс ускорителя”, В. Якимчук, техническое описание.

[15] “PCI-7200 Datasheet”, http://www.adlinktech.com

[16] “PCI-7200 / cPCI-7200 12MB/S High Speed Digital Input/ Output Card”, PCI7200 manual, Copyright 1999 ADLink Technology Inc.

[17] “CY7C68001 EZ-USB SX2 High Speed USB Interface Device datasheet”, Cypress Semiconductor Corporation, 2004.

[18] “Universal Serial Bus Revision 2.0 specification”, http://www.usb.org

[19] “ALICE DAQ DATE the data acquisition software”, Filippo Costa on behalf of the ALICE DAQ GROUP, ALICE Upgrade Workshop, March 2012.

[20] “Аппаратура для подключения электронных систем МИСС, КАМАК и СУММА к персональному компьютеру”, Петров В.С., Якимчук В.И., Препринт ИФВЭ 2011-21, Протвино, 2011.

[21] “Parallel Virtual File System, Version 2”, PVFS2 Development Team, PVFS Developer’s Guide, http://www.pvfs.org

[22] “Silent corruptions.”, Kelemen, 2007, In 8th Annual Workshop on Linux Clusters for Super Computing.

[23] “A Conversation with Jeff Bonwick and Bill Moore”, Association for Computing Machinery, November 15, 2007, http://queue.acm.org/detail.cfm?id=1317400

[24] “GlusterFS is a software-only, highly available, scalable, centrally managed storage pool for public and private cloud environments.”, домашняя страница GlusterFS, http://www.gluster.org

[25] “Open Source High-Availability Software for Linux and other Platforms”, домашняя страница проекта програмного обеспечения для построения высоконадёжных систем на основе Linux и других платформ, http://linux-ha.org/wiki/Split_Brain

[26] “OpenFabrics Alliance overview”, домашняя страница OpenFabrics Alliance, https://www.openfabrics.org/home/ofa-overview.html

[27] “Infiniband, 10GigE and GlusterFS.”, обзор транспортных протоколов от разработчика GlusterFS Anand Babu, http://www.unlocksmith.org/2009/11/infiniband-10gige-and-glusterfs.html

[28] “Cloud Storage for the Modern Data Center: An Introduction to Gluster Architecture Versions 3.1.x”, Copyright 2011, Gluster, Inc.

[29] “Неуправляемый гигабитный коммутатор 2 уровня с 24 портами 10/100/1000Base-T и встроенным источником питания DGS-1024D”, http://www.dlink.ru

[30] “Universal Serial Bus 3.0 Specification”, http://www.usb.org

[31] “EZ-USB FX3 SuperSpeed USB Controller datasheet”, Cypress Semiconductor Corporation, 2012.

[32] “Bonding”, The Linux Foundation, http://www.linuxfoundation.org

[33] “IEEE 802.3ab”, IEEE Standarts Association, http://standards.ieee.org

[34] “FPGA based data acquisition system for COMPASS experiment”, M. Bodlak, V. Frolov, V. Jary, S. Huber, I. Konorov, D. Levit, J. Novy, S. Paul, R. Salac and M. Virius, October 2013, arXiv:1310.1308v1 [physics.ins-det]

*Рукопись поступила 4 декабря 2013 г.*

Препринт отпечатан с оригинала-макета, подготовленного авторами.

С.В. Донсков и др.
Система сбора данных эксперимента ОКА.

Оригинал-макет подготовлен с помощью системы **LaTeX.**

---

Подписано к печати 11.12.2013. Формат 60 × 84/16. Цифровая печать.
Печ.л. 2. Уч.-изд.л. 2,88. Тираж 80. Заказ 51. Индекс 3649.

---

ФГБУ ГНЦ ИФВЭ
142281, Протвино Московской обл.

Индекс 3649

П Р Е П Р И Н Т 2013–22, И Ф В Э, 2013