УДК 519.95

«Создание хранилищ данных большого объема. Практический опыт»

Белоконный А.В.

«Creating The Data Warehouse of Huge Information Content. Practical experience»

Belokonnyy A.

Аннотация

В статье обсуждается практика разработки и внедрения вычислительных комплексов для систем обработки и хранения больших объемов данных на базе мейнфреймов и картриджных библиотек. В качестве примера рассматривается издание хранилища данных на 5 Пбайт и технологии работы с ним.

Abstracts

The article discusses the practice of developing and implementing computer systems for data processing and storage of large amounts of data based on the mainframe and cartridge libraries. As an example, the publication of the data warehouse at 5 petabytes, and technology to work with.

Ключевые слова: Хранилище информации, Мейнфрейм, Картриджная библиотека, Иерархическая файловая система.

Data Warehouse: Mainframe, Cartridge library, Hierarchical File System

В ОАО «НИЦЭВТ» разрабатываются и выпускаются вычислительные комплексы для хранения и обработки больших объемов информации.

В основе этих комплексов лежат

- одна или несколько вычислительных машин большой производительности;
- один или несколько дисковых массивов для оперативного хранения информации;
- библиотеки для длительного хранения информации;
- сертифицированное программное обеспечение разработки ОАО «НИЦЭВТ» и фирменное программное обеспечение IBM;
- средства инженерного обеспечения, в том числе бесперебойное питание и отвод тепла.

При разработке вычислительных комплексов необходимо принимать во внимание многие факторы, в том числе:

- масштаб решаемых задач - по количеству, мощности, регламенту;
- категорийность решаемых задач и технология обработки задач составляющих государственную или коммерческую тайны;
- соотношение между объемами оперативной информации и информации длительного хранения;
- ограничительные факторы – площадь, потребляемые мощности, время восстановления процесса после сбоя или отказа технических или программных средств, стоимостные характеристики решений.

В докладе кратко обсуждается практика разработки и внедрения вычислительных комплексов для систем обработки и хранения больших объемов данных.

В качестве основной машины вычислительного комплекса выступает машина IBM zSeries отличительными особенностями которой являются – большое количество процессоров в одной коробке - до 96-и, каждый процессор тактовой частотой 5,3 ГГц, до 192 каналов ввода/вывода, потребление электроэнергии 5-10 кВА в зависимости от кол-ва открытых процессоров.

Дисковые подсистемы – от различных производителей (IBM, Hitachi,….) Могут одновременно функционировать как в открытых (FiberChannel) сетях, так и мейнфреймных сетях (FICON). Объем от 5-100 ТБ на одну стойку. Потребление электроэнергии от 3,5 до 15 кВА.

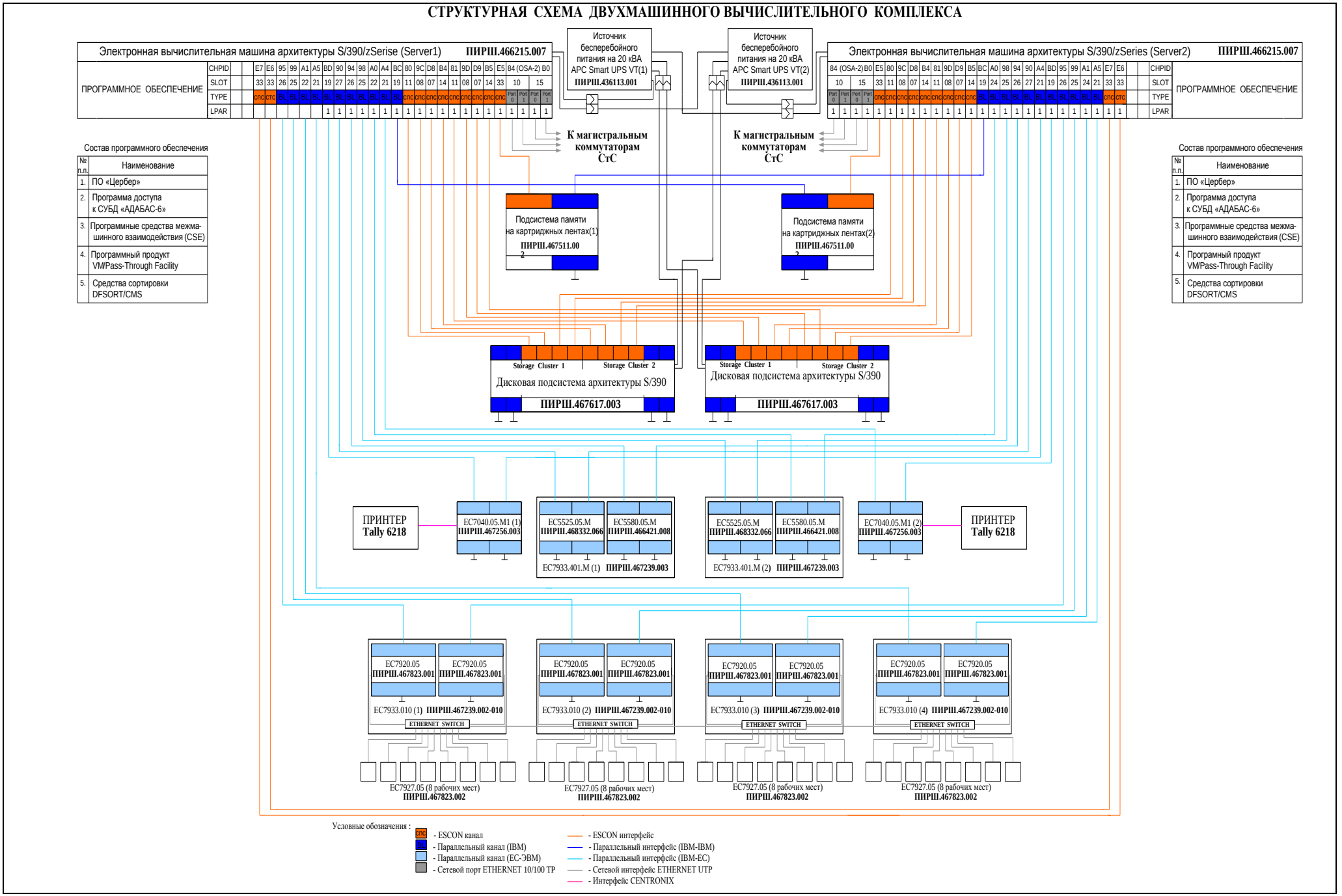


Рис. 1 Типовой двухмашинных, отказоустойчивый вычислительный комплекс, поставляемый ОАО «НИЦЭВТ»

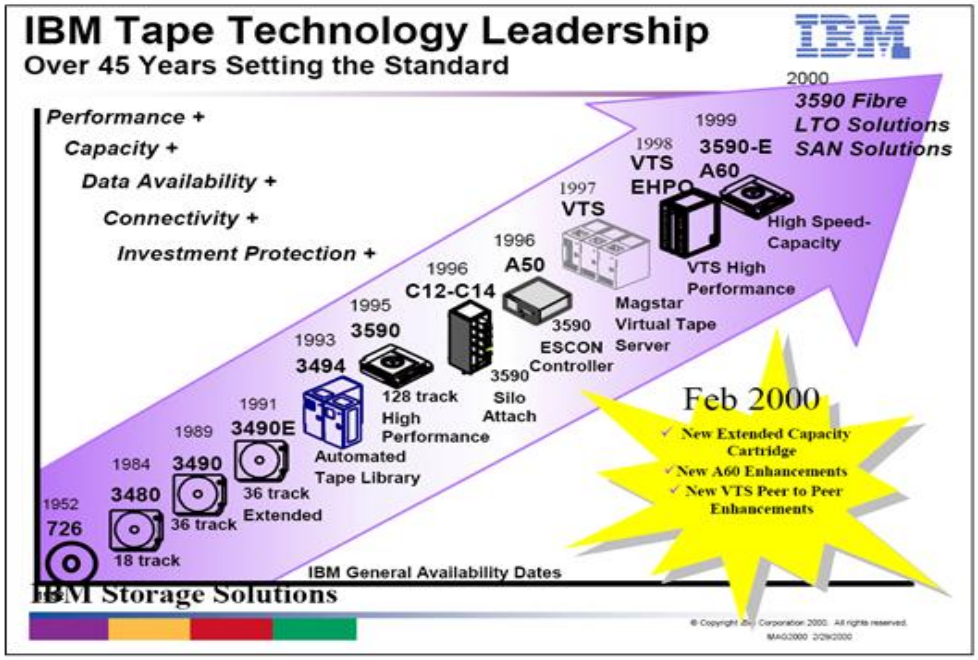


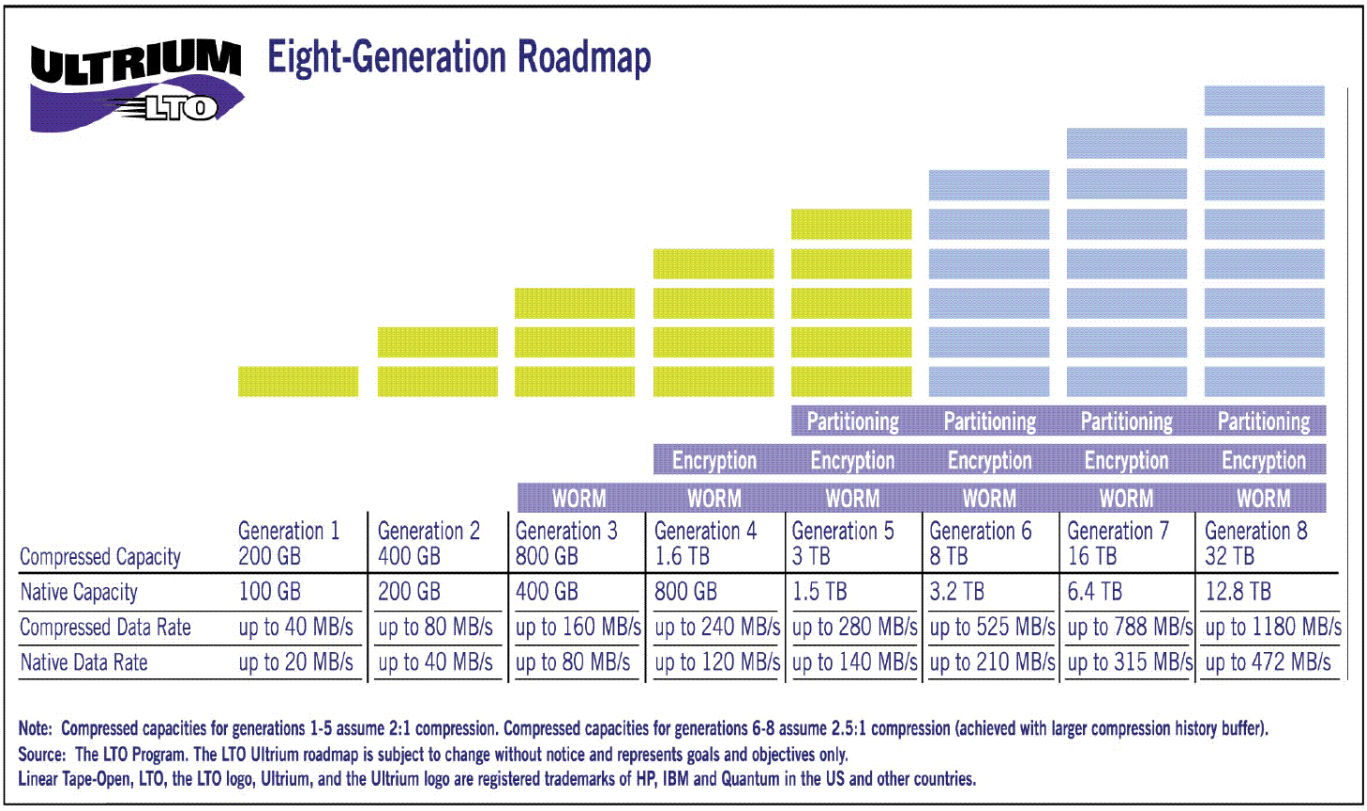
Рис. 2 Краткая история развития ленточных и картриджных систем фирмы IBM с 1952 по 2000 годы.

IBM System Storage TS1120 Tape Drive 	Емкость картриджа	0,7 - 2,1 ТБ (при компрессии 3:1)
	Скорость обмена информацией	104 МБ/сек для несжатой информации
	Интерфейс	ESCON/FICON/FC
	Тип картриджа	3592
	Год выпуска	2006
IBM System Storage TS1130 Tape Drive 	Емкость картриджа	1,0 – 3,0 ТБ (при компрессии 3:1)
	Скорость обмена информацией	160 МБ/сек для несжатой информации
	Интерфейс	ESCON/FICON/FC
	Тип картриджа	3592
	Год выпуска	2009
IBM System Storage TS1140 Tape Drive 	Емкость картриджа	4,0 – 12,0 ТБ (при компрессии 3:1)
	Скорость обмена информацией	250 МБ/сек для несжатой информации
	Интерфейс	FICON/FC
	Тип картриджа	**3592**
	Год выпуска	**2011**

Рис.3 характеристики современных картриджных устройств чтения/записи информации.

Картриджные библиотеки. Объем хранения от 24 – 3,000 ТБайт на стойку (шкаф), в случае наращивания дополнительными стойками сотни ПБайт. Потребление картриджной библиотеки в районе 1,5 кВА и определяется не количеством шкафов, а контроллером управления, количеством роботов и устройств чтения-записи.

На рис. 1 в качестве примера приведена реализация двухмашинного отказоустойчивого комплекса, включающего вычислительные машины, дисковые и картриджные подсистемы хранения, а также терминальные устройства. Операционные системы, средства отказоустойчивости, СУБД (Система управления базами данных) и СУРД (Система управления разграничением доступа) сертифицированы на обработку категорированной информации, составляющей Государственную тайну. Комплекс функционирует в режиме горячего резервирования.



Compressed Capacity	Generation 1 200 GB	Generation 2 400 GB	Generation 3 800 GB	Generation 4 1.6 TB	Generation 5 3 TB	Generation 6 8 TB	Generation 7 16 TB	Generation 8 32 TB
Native Capacity	100 GB	200 GB	400 GB	800 GB	1.5 TB	3.2 TB	6.4 TB	12.8 TB
Compressed Data Rate	up to 40 MB/s	up to 80 MB/s	up to 160 MB/s	up to 240 MB/s	up to 280 MB/s	up to 525 MB/s	up to 788 MB/s	up to 1180 MB/s
Native Data Rate	up to 20 MB/s	up to 40 MB/s	up to 80 MB/s	up to 120 MB/s	up to 140 MB/s	up to 210 MB/s	up to 315 MB/s	up to 472 MB/s

Note: Compressed capacities for generations 1-5 assume 2:1 compression. Compressed capacities for generations 6-8 assume 2.5:1 compression (achieved with larger compression history buffer).
Source: The LTO Program. The LTO Ultrium roadmap is subject to change without notice and represents goals and objectives only.
Linear Tape-Open, LTO, the LTO logo, Ultrium, and the Ultrium logo are registered trademarks of HP, IBM and Quantum in the US and other countries.

Рис. 4. Перспективы развития LTO технологии

В качестве основного хранилища данных в вычислительных комплексах производства используются картриджные библиотеки.

Выбор картриджных библиотек в качестве хранилищ данных большого объема вызван рядом факторов:

- Большой объем хранения на единицу площади;
- Низкий уровень потребления электроэнергии;
- Высокая надежность хранения информации, компенсация сложных ошибок, в том числе физического повреждения носителя;
- Гарантированное производителем хранение информации в течение 30 лет;
- Минимальная стоимость хранения на 1 ТБ – на современных устройствах она составляет 60 USD/ТБ без сжатия информации;

Ленточные системы в своем развитие прошли долгую историю:

На рис.2 приведена история развития ленточных и картриджных носителей фирмы IBM с 1952 по 2000 год, а на рис. 3 представлены характеристики современных картриджных устройств чтения/записи информации.

Особенностью этих устройств является то, что они используют один и тот же тип носителя - картридж IBM 3592. Увеличения емкости хранения достигается совершенствованием технологии чтения/записи.

Современный картридж позволяет хранить до 4/12 ТБ информации (несжатом/ сжатом виде), при стоимости носителя порядка 4,000 рублей.

По типу записи картриджи 3592 относятся к усовершенствованным LTO.

На рисунке 4 приведены перспективы развития технологии LTO.Современное состояние технологии – это Generation 6 – 7. Ближайшие перспективы – это по крайней мере 12,8 ТБ на 1 картридж без сжатия.

При создании хранилищ данных большого объема в рамках вычислительных комплексов, в том числе с использованием разнородных системотехнических платформ решаются следующие основные задачи:

1) Проведение резервного копирования/восстановления (Backup/Restore) для мейнфреймов.

2) Проведение резервного копирования/восстановления для открытых систем (UNIX, LINUX, WINDOWS, ORACLE и т.д.).

3) Создание иерархической файловой системы хранилища данных на базе картриджных библиотек доступных всем пользователям комплекса в рамках их полномочий.

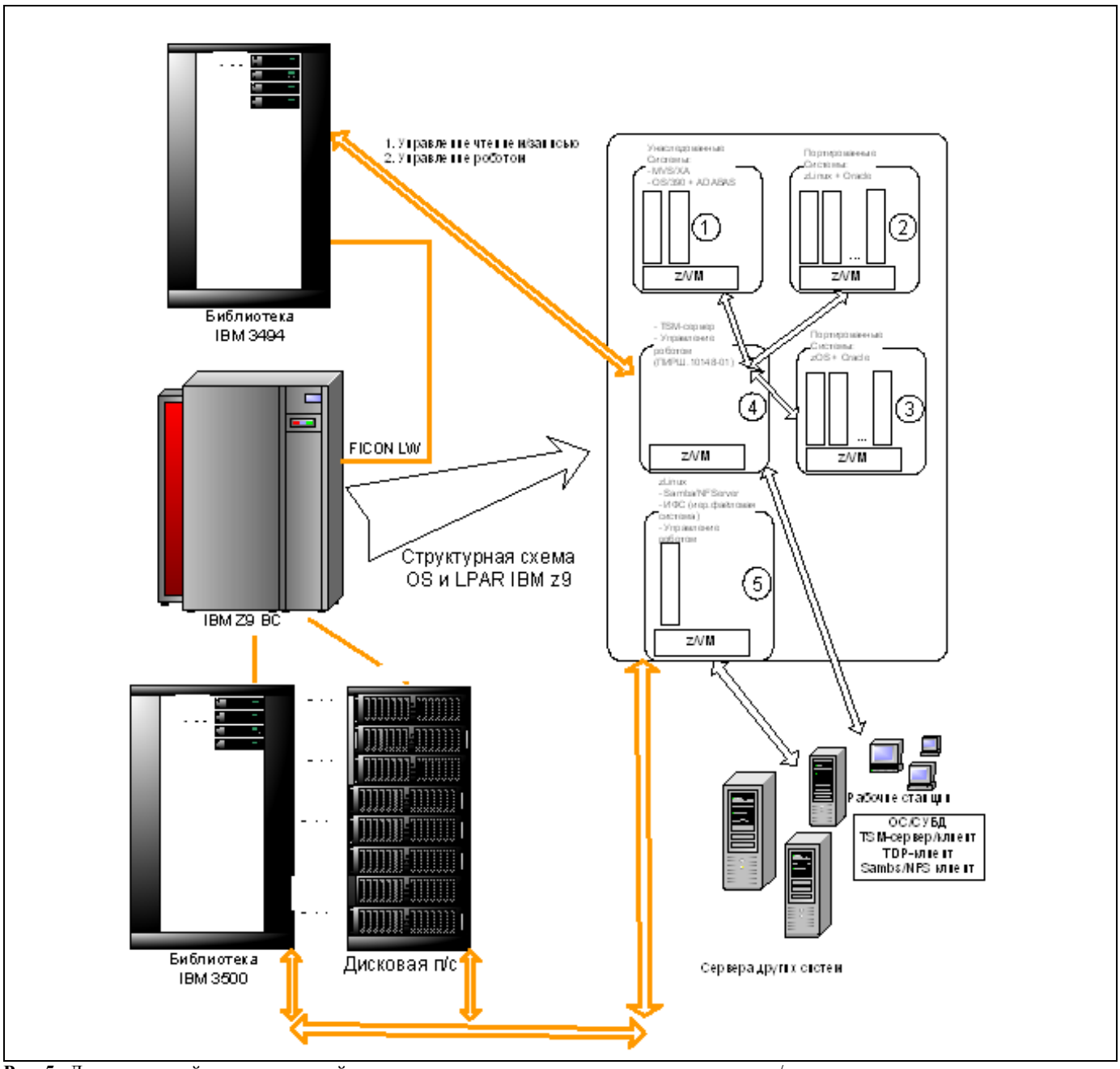


Рис. 5. Двухмашинный вычислительный комплекс с решением задач резервного сохранения/восстановления и создания иерархической файловой системы.

На рисунке приведен пример реализации двухмашинного вычислительного комплекса с решением задач резервного сохранения/восстановления и создания иерархической файловой системы.

Реализуется с помощью разработанных в ОАО «НИЦЭВТ» программных продуктов:

- «Система файловая иерархическая ПИРШ.10147-01», позволяющая размещать наборы данных, как на дисковых подсистемах, так и на картриджах.
- «Система библиотечная иерархическая ПИРШ.10148-01» для управления роботом библиотеки.

Использование, разработанных в ОАО «НИЦЭВТ» программных продуктов позволяет создать доверенную (сертифицируемую) программную среду при формировании больших хранилищ данных.

Резервное копирование и восстановление данных:

На рисунке 6 приведен состав архивируемых/восстанавливаемых томов вычислительного комплекса, их размещение на различных физических носителях и типы подключения этих носителей к вычислительной машине.

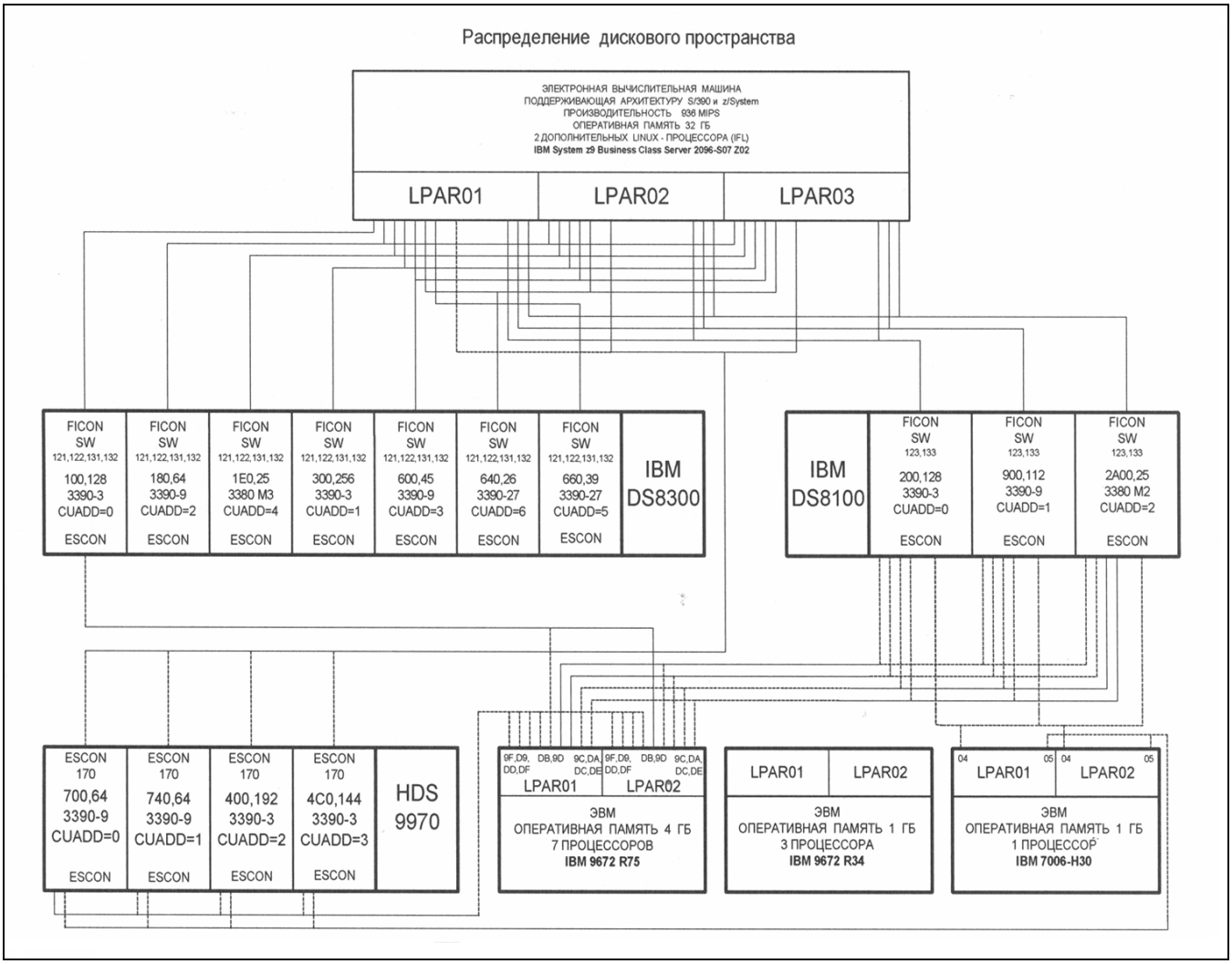


Рис.6. Состав архивируемых томов вычислительного комплекса

Рис.7. Масштабируемость картриджных библиотек.

Особенность реализации состоит в том, что в составе вычислительного комплекса присутствуют устаревшие носители информации, подключаемые к вычислительной машине по каналам ESCON (16 Мбайт/сек) и современные носители с подключением по FICON-каналам (1/2/4 Гбит/сек).

Время архивации/восстановления томов данных сильно зависит от типа носителя, на которых том размещен или должен быть восстановлен и типа подключения к вычислительной машине.

Достигнуты следующие результаты при проведении полного и инкрементального резервного копирования:

Полное резервное копирование - 724 тома общей емкостью 500 ГБ из них 70 ГБ ORACLE с SUN продолжительность (20 часов).

Инкрементальное резервное копирование порядка 200 томов – 3 часа.

Столь длительное время резервного копирования объясняется в первую очередь наличием большого количества 100/200 Мбайтных томов, расположенных на носителях, подключаемых к вычислительной машине по ESCON (медленным каналам).

Во время резервного копирования включено время переноса данных с ORACLE/SUN по сети Ethernet на дисковую подсистему DS8300 (порядка 3-х часов).

Создание иерархической системы хранения

На технических средствах того же вычислительного комплекса было создано иерархическое файловое хранилище объемом – 10 ТБайт на дисковых подсистемах с расширением на 5 Пбайт на картриджной системе. Иерархическая файловая система была доступна пользователям через NFS-интерфейс в сети.

В качестве библиотеки использовалась библиотека IBM 3494 в одностоечном варианте с 4-я устройствами чтения/записи IBM TS1130.

Порядок функционирования иерархической файловой системы:

1) Запрос на доступ к файлу от пользователя – вызывает процедуру открытия файла - open(file) – после этого возможны два варианта - либо данные на диске, либо имеется только ссылка на эти данные;

2) В случае отсутствия файла на диске выполняется процедура разрешения ссылки, как адрес данных в библиотеке (№ картриджа, данные из оглавления картриджа);

3) Роботом библиотеки выполняется установка картриджа в устройство чтения/записи(2 сек).

4) Устройство чтения/записи выполняет чтение оглавления картриджа в устройство чтения/записи (зависит от объема оглавления) – в обычных ситуациях секунды;

5) Поиск данных на ленте – до 60 сек. в худшем случае;

6) Выполняется пересылка данных на диск.

В итоге процедура доступа на чтение к данным, размещенным на картриджной системе, не превышает одной минуты. При этом доступно для формирования хранилища информации 5 Пбайт.

Выводы:

Большие хранилища данных на базе картриджных библиотек программного обеспечения разработки ОАО «НИЦЭВТ» позволяют:

1) Создавать масштабируемые хранилища данных от десятков ТБ до сотен ПБ простым наращиванием;

2) Обеспечение сохранения инвестиций – каждые три года происходит увеличение объемов хранения без смены носителя;

3) Обеспечить единую среду хранения информации для мейнфреймов и открытых систем (UNIX, LINUX, WINDOWS);

4) Обеспечить минимальное потребление энергии на объем хранения;

5) Создавать большие хранилища данных для обработки и хранения категорированной информации, составляющей Государственную тайну.