OPEN SYSTEMS STORAGE PLATFORMS

by

KIRBY COLLINS
CONVEX STORAGE SYSTEMS

Open Systems
Storage Platforms

| Storage Applications |
| --- |

ConvexOS System Software

High Performance Peripherals

DS Series Data Server Systems

The building blocks for an open storage system includes a system platform, a selection of storage devices and interfaces, system software, and storage applications

CONVEX storage systems are based on the DS Series Data Server systems. These systems are a variant of the C3200 supercomputer with expanded I/O capabilities. These systems support a variety of medium and high speed interfaces to networks and peripherals. System software is provided in the form of ConvexOS, a POSIX compliant derivative of 4.3BSD UNIX. Storage applications include products such as UNITREE and EMASS.
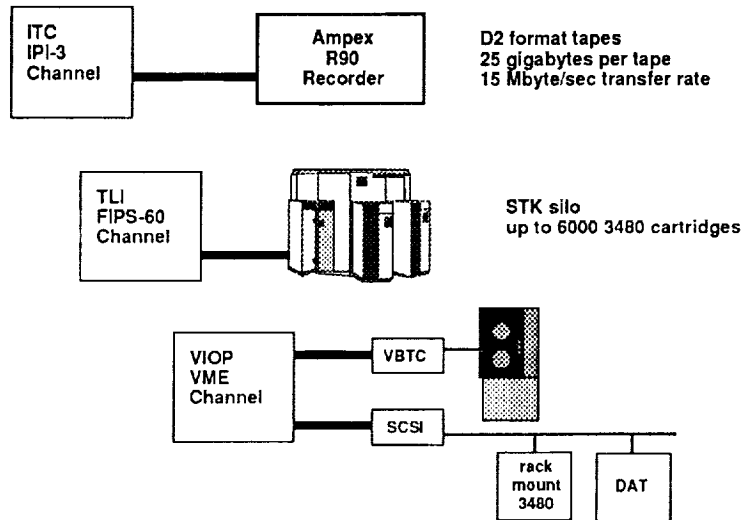
DS Series Systems
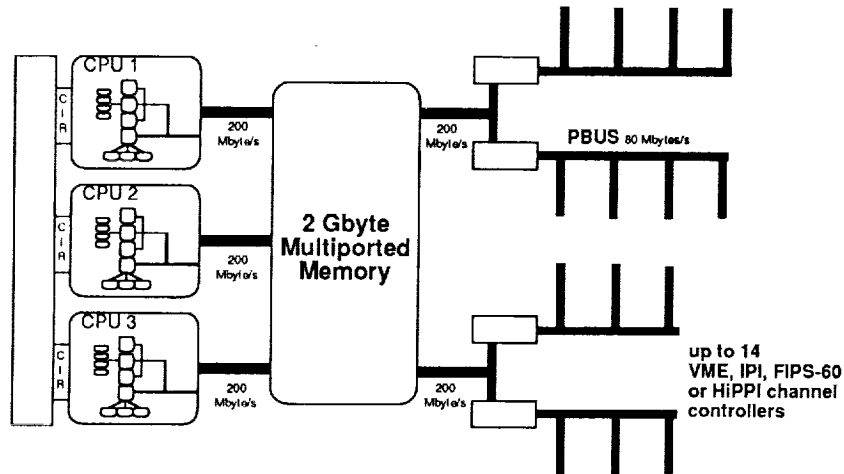
The performance of the DS Series is driven by the main memory system, a multiported arrangement with up to 2 Gigabytes of RAM and a total bandwidth of 800 megabytes/second. A crossbar connects this memory system to five 64bit wide ports, each capable of transferring data at up to 200 megabytes/second. In a C3200 supercomputer, four of these ports are dedicated to CPUs, with the fifth attached to the I/O system. In the DS Series version of the architecture, two memory ports are allotted to I/O and only three CPUs are supported.

The I/O system is designed around a high speed I/O bus called the PBUS. Each memory port supports two PBUS's, for a total of four in the DS Series. The PBUS in turn supports intelligent channels, each of which supports one or more VME, IPI, FIPS-60, or other types of standard I/O busses.

Peripherals - Tape Systems

```
┌──────────┐          ┌──────────┐
│ ITC      │          │ Ampex    │       D2 format tapes
│ IPI-3    │──────────│ R90      │       25 gigabytes per tape
│ Channel  │          │ Recorder │       15 Mbyte/sec transfer rate
└──────────┘          └──────────┘

┌──────────┐
│ TLI      │
│ FIPS-60  │──────────              STK silo
│ Channel  │                        up to 6000 3480 cartridges
└──────────┘

┌──────────┐      ┌──────┐
│ VIOP     │──────│ VBTC │─────
│ VME      │      └──────┘
│ Channel  │      ┌──────┐
└──────────┘──────│ SCSI │─────
                  └──────┘
                        ┌───────┐   ┌─────┐
                        │ rack  │   │ DAT │
                        │ mount │   │     │
                        │ 3480  │   │     │
                        └───────┘   └─────┘
```
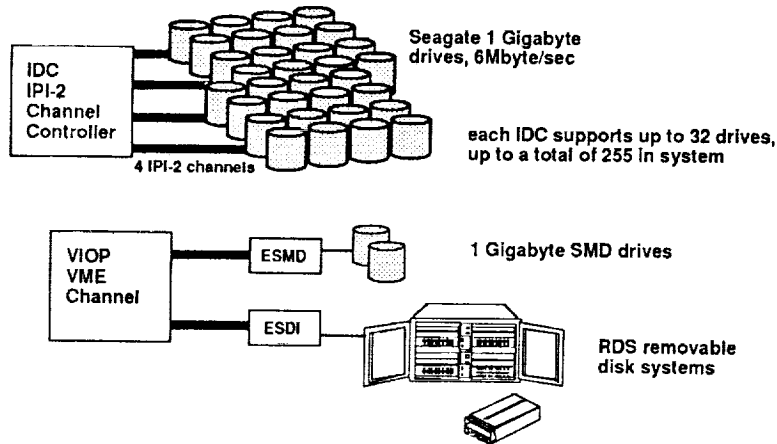
A range of tape systems is available as well - again based on industry standard interfaces.

A channel controller will soon be available which generates four IPI-3 channels, which can be used to interface such devices as the Ampex R90 DD2 recorder. This device uses a tape format based on video broadcast technology to store from 25 to 150 gigabytes per tape, depending on cartridge size, with transfer rates of up to 15 gigabytes/second.

The TLI generates two FIPS-60 (IBM Block Mux Channel) interfaces, and can be used to connect to such devices as the Storage Technologies ACS (popularly known as the silo) or high duty cycle 3480 compatible tape transports.

Lastly, VME cards can be used to interface to lower speed devices such as nine track tapes, or SCSI based devices such as DAT or medium duty cycle rack mount 3480 compatible drives.

Peripherals - Disk Systems

IDC IPI-2 Channel Controller

4 IPI-2 channels

Seagate 1 Gigabyte drives, 6Mbyte/sec

each IDC supports up to 32 drives, up to a total of 255 in system

VIOP VME Channel

ESMD

1 Gigabyte SMD drives
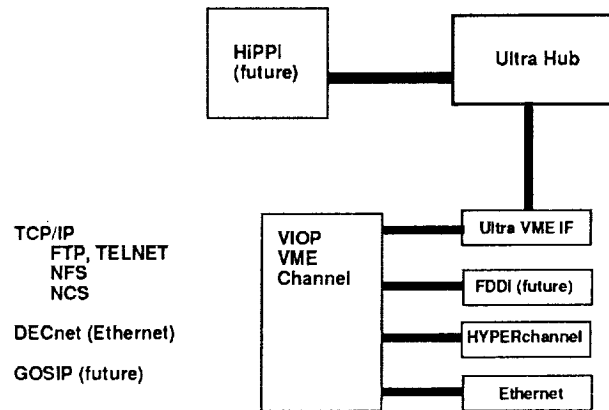
ESDI

RDS removable disk systems

CONVEX supports three basic kinds of disk systems.

High performance IPI-2 disk drives are supported via the IDC, an 88000 based integrated channel controller which generates four IPI-2 interfaces, each capable of transferring data at up to 10 megabytes/second. Each of those interfaces can support up to 8 disk drives, for a total of 32 per IDC. With multiple IDC's more drives can be attached, up to the system maximum of 255. The system is capable of transferring data into these drives at aggregrate rates in excess of 50 megabytes per second.

Lower speed drives are supported via the VIOP VME channel, which supports up to two VME busses with 10 megabytes/second of bandwidth each. An SMD-E interface is available along with 1 gigabyte winchester drives, and an ESDI interface is also supported for low cost 780 megabyte drives.
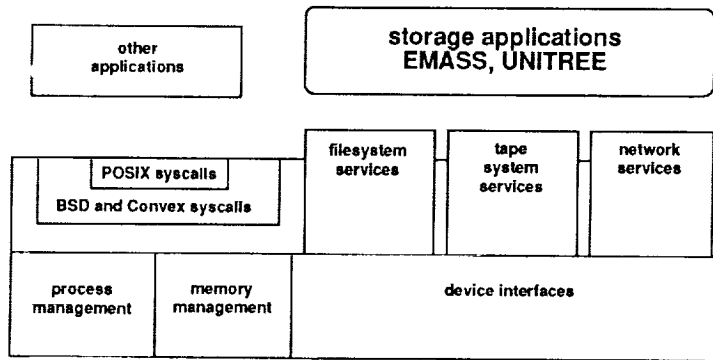
Peripherals - Networks

```
        ┌──────────┐              ┌──────────┐
        │ HiPPI    │══════════════│ Ultra Hub│
        │ (future) │              │          │
        └──────────┘              └──────────┘
                                        ║
                                        ║
  TCP/IP                     ┌────────┐  ┌──────────────┐
     FTP, TELNET             │ VIOP   │══│ Ultra VME IF │
     NFS                     │ VME    │  └──────────────┘
     NCS                     │ Channel│  ┌──────────────┐
                             │        │══│ FDDI (future)│
  DECnet (Ethernet)          │        │  └──────────────┘
                             │        │  ┌──────────────┐
  GOSIP (future)             │        │══│ HYPERchannel │
                             │        │  └──────────────┘
                             │        │  ┌──────────────┐
                             │        │══│ Ethernet     │
                             └────────┘  └──────────────┘
```

Network interfaces are provided primarily through VME based interface cards. Controllers are available for Ethernet, HYPERchannel, and Ultra, with FDDI available in the near future. A HiPPI channel is also planned in the near future, to interface to the Ultra hub at high bandwidths, as well as other HiPPI based devices and switches.
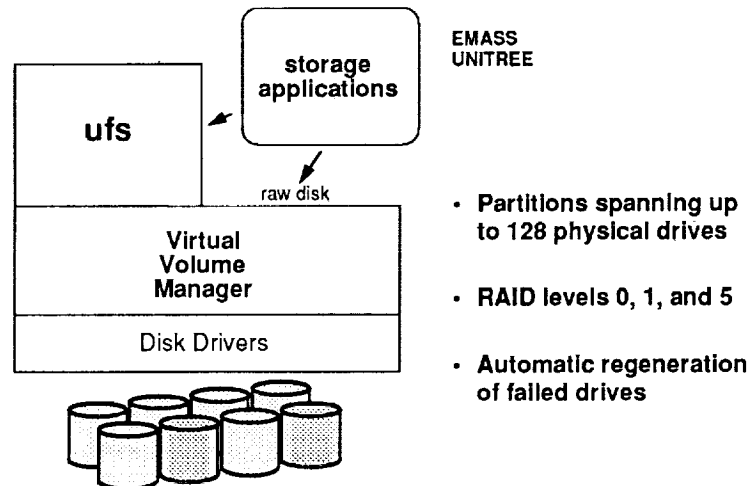
The operating system supports the most commonly used protocols, such as TCP/IP, DECnet (Ethernet only), and GOSIP (in the future).

ConvexOS

```
┌─────────────────┐    ┌──────────────────────────┐
│     other       │    │   storage applications   │
│  applications   │    │     EMASS, UNITREE       │
└─────────────────┘    └──────────────────────────┘

┌─────────────────────┐ ┌──────────┬──────────┬──────────┐
│  ┌───────────────┐  │ │filesystem│   tape   │ network  │
│  │ POSIX syscalls│  │ │ services │  system  │ services │
│  └───────────────┘  │ │          │ services │          │
│ BSD and Convex syscalls│         │          │          │
├──────────┬──────────┤ ┴──────────┴──────────┴──────────┤
│ process  │ memory   │      device interfaces           │
│management │management│                                  │
└──────────┴──────────┴──────────────────────────────────┘
```

The heart of any open systems strategy is the system software, since it is compatibility at the system interface that allows a high degree of interoperability and portability of storage applications.

ConvexOS is a derivative of 4.2BSD UNIX, updated with 4.3BSD enhancements, and modified to be compliant with the POSIX.1 standard. Semaphored for symmetric multiprocessing, it uses the parallel architecture of the C3200/DS architecture to provide high throughput under a heavy I/O load. Extensions to the basic UNIX filesystem and tape services provide support for storage applications such as EMASS and UNITREE.
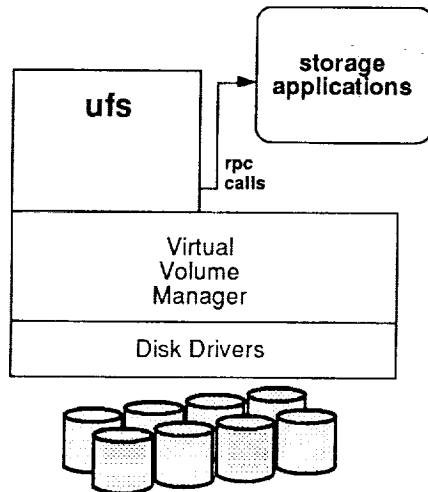
Disk Management

	storage applications	EMASS UNITREE
ufs		

raw disk

Virtual Volume Manager

Disk Drivers

- Partitions spanning up to 128 physical drives

- RAID levels 0, 1, and 5

- Automatic regeneration of failed drives

One clear trend in storage is towards large numbers of magnetic disks. To ensure that the connection of large numbers of disk drives does not reduce the overall reliability of a storage system, ConvexOS provides a driver layer called the Virtual Volume Manager (VVM).

VVM takes up to 128 physical partitions on different disks, and combines them into one large partition to the file system and other higher level applications. VVM interleaves successive blocks across different drives to allow parallel access for sequential disk I/O, and can generate mirror or parity blocks so that a failure of any one disk does not cause loss of data. If a drive does fail, VVM can reconstruct data blocks on a spare drive,so that the system can tolerate a drive failure and return to redundant operation, all without interrupting applications that are accessing the partition or requiring operator intervention.

File Fault Interface

```
                    ┌──────────────┐
                    │   storage    │
┌─────────────┐     │ applications │
│             │────▶│              │
│     ufs     │     └──────────────┘
│             │      rpc
│             │      calls
│             ├─────┐
├─────────────┤     │
│   Virtual   │     │
│   Volume    │     │
│   Manager   │     │
├─────────────┤─────┘
│ Disk Drivers│
└─────────────┘
```

- filesystem events can trigger rpc calls to an application
  - file open, read/write
  - metadata update
  - block allocation
  - space watermarks
- allows storage applications to manipulate files transparently to user applications

One of the most difficult challenges in supporting a variety of storage applications is how to tightly integrate these applications with the host system's own file system. Since the host filesystem is usually highly tuned to the architecture, it may make sense to use it for local buffering of files, especially if the host is also used for compute intensive applications as well as storage. Many storage applications have client code that is inserted in the kernel of the host operating system. Providing support for client code for very many storage applications is problematic, since there is little uniformity in these interfaces.

To resolve this ConvexOS provides an interface which allows a storage application to trap many filesystem events, with notification via RPC calls to a daemon process outside the kernel. This allows a storage application to be notified of file creation and access, and migrate data blocks into and out of the filesystem on demand. This activity is completely transparent to applications accessing files through the ConvexOS filesystem, and allows hierarchical storage management to be implemented without modification to applications.

Convex Storage Servers

- Assembled from off-the shelf components
- Industry standard interfaces
- Convex engineering for performance and reliability

With the DS Series of storage systems, Convex has developed a set of products which provide open systems solutions for storage management applications. The systems are highly modular, assembled from off the shelf components with industry standard interfaces. The C Series system architecture provides a stable base, with the performance and reliability of a general purpose platform.

This combination of a proven system architecture with a variety of choices in peripherals and application software allows wide flexibility in configurations, and delivers the benefits of open systems to the mass storage world.