

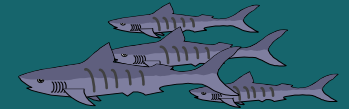
IBM Almaden Research Center

Tiger Shark

*A scalable, fault-tolerant file system for
video-on-demand.*

*Roger Haskin
IBM Almaden Research Center
San Jose, California*

Tiger Shark Applications



IBM Almaden Research Center

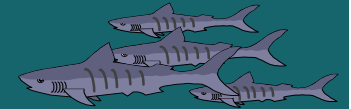
- *Video on Demand and Interactive TV*
 - *Bell Atlantic Field Trial - 50 MPEG1 streams (4/93-9/94)*
 - *Hong Kong Telecom Trial - 150 MPEG1 streams (2/95-9/95)*
 - *Tokyo Metropolitan Govt. VOD Trial - 100 MPEG2 (mid-'96)*
- *LAN-based Multimedia*
 - *IBM Multimedia LAN Server for AIX - real-time delivery to NFS clients over switched Ethernet, FDDI, or ATM*
 - *Technology in OS/2 LAN Server Ultimedia*
- *Video over the Internet*
 - *Supercomputing '95 - MJPEG video from SP-2 to modified VIC clients via RTP over ATM WAN (M-Bone) - done in conjunction with Argonne Labs*

Video Delivery System Components

IBM Almaden Research Center

- *Client (PC or set-top box)*
 - *Controls flow of video (start, stop, seek, ...)*
 - *Presents video to user (decode, display)*
- *Network*
 - *Physical media (T1, ATM, LAN, ...)*
 - *Protocol driver (AAL5, RTP, NFS, ..., or none)*
- *File System*
 - *High capacity*
 - *High throughput*
 - *Real-time delivery*

Why a Special File System?

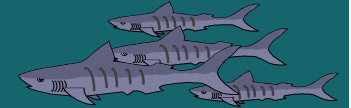


IBM Almaden Research Center

With a conventional file system.....

- *No real-time capability*
 - *video/audio compete with non-realtime I/O*
 - *scheduling not appropriate (e.g. elevator)*
- *Poor disk performance*
 - *File system optimized for space, not throughput*
- *Limited scale-up*
 - *Capacity limitations (e.g. 2GB file size).*
 - *Growth limitations (e.g. single machine)*

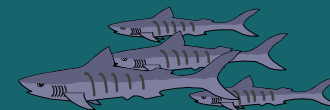
Tiger Shark File System



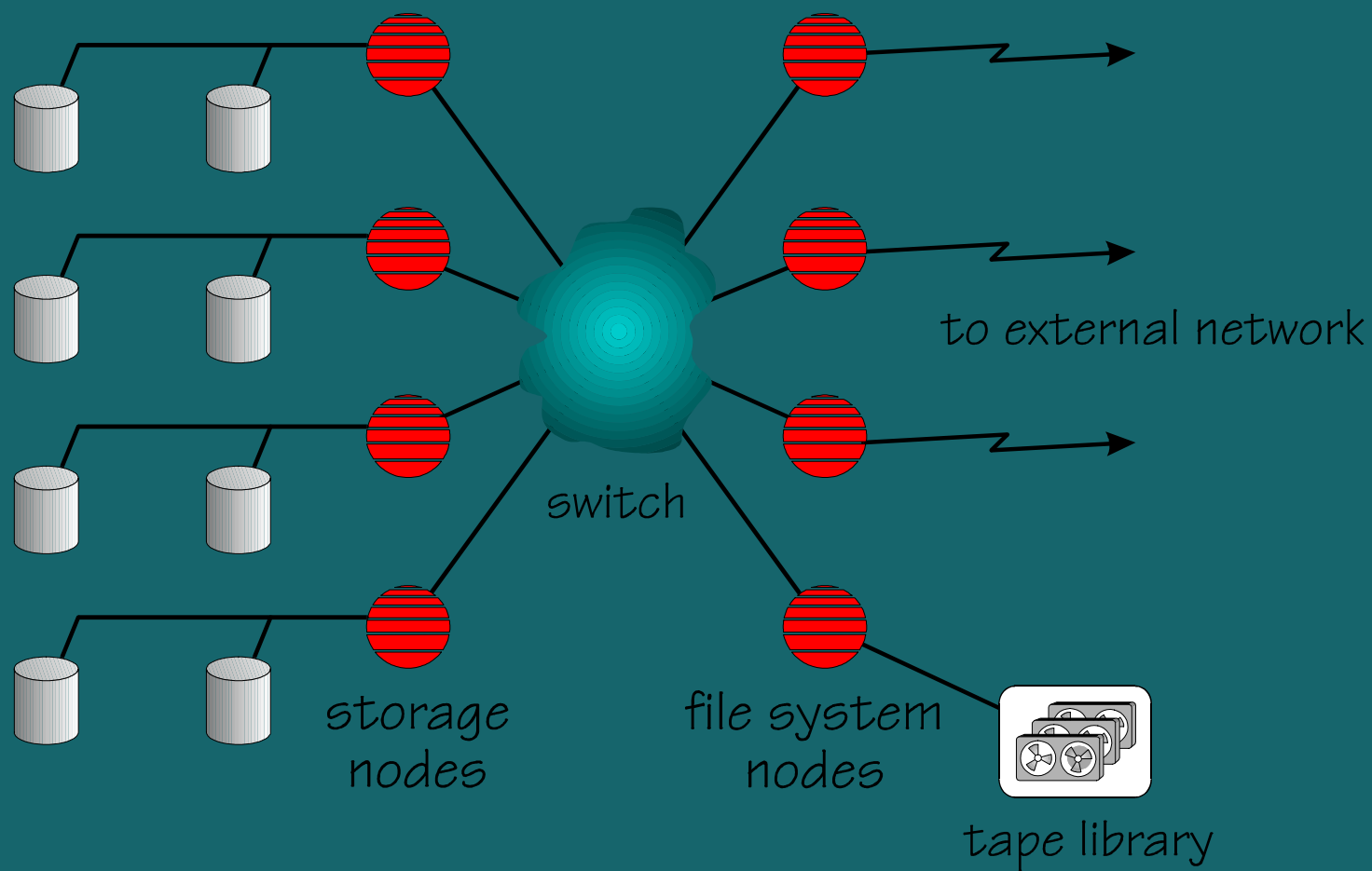
IBM Almaden Research Center

- *Parallel, shared-disk* file system for IBM SP-2 supports cache-coherent multinode read/write sharing
- *Wide striping* for high bandwidth, scalability, and automatic load balancing
- *Quality of Service* features include admission control, real-time disk scheduling
- *Fault-tolerance* - data replication, log-based recovery
- *System management* commands support online reconfiguration
- *Posix-compliant* to support conventional applications

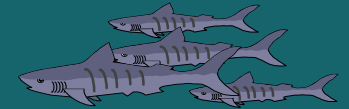
Tiger Shark on SP-2



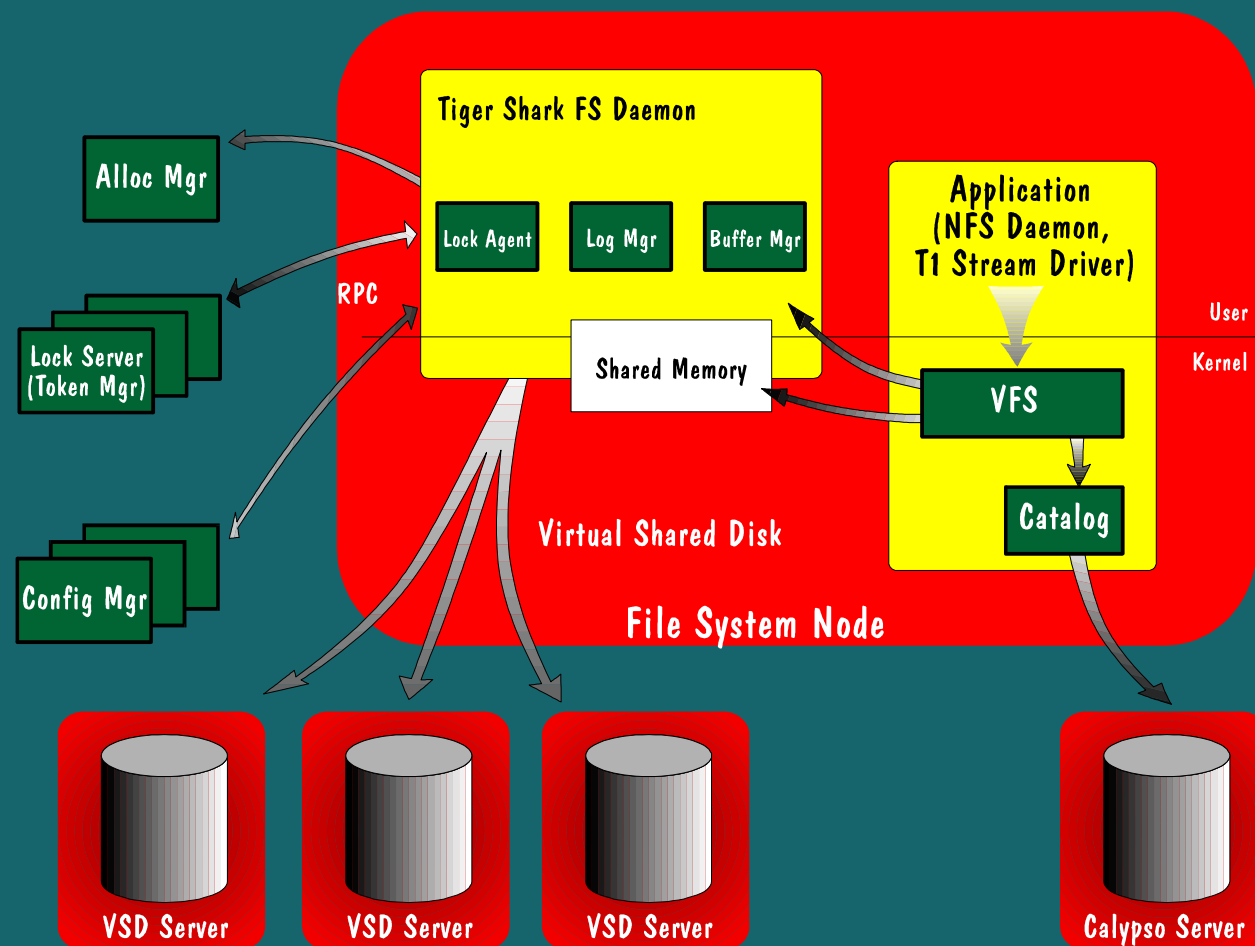
IBM Almaden Research Center



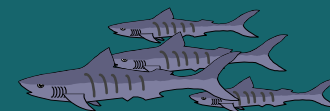
Tiger Shark Software Structure



IBM Almaden Research Center



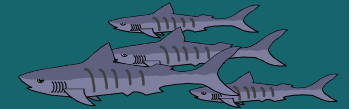
Real-Time Disk Scheduling



IBM Almaden Research Center

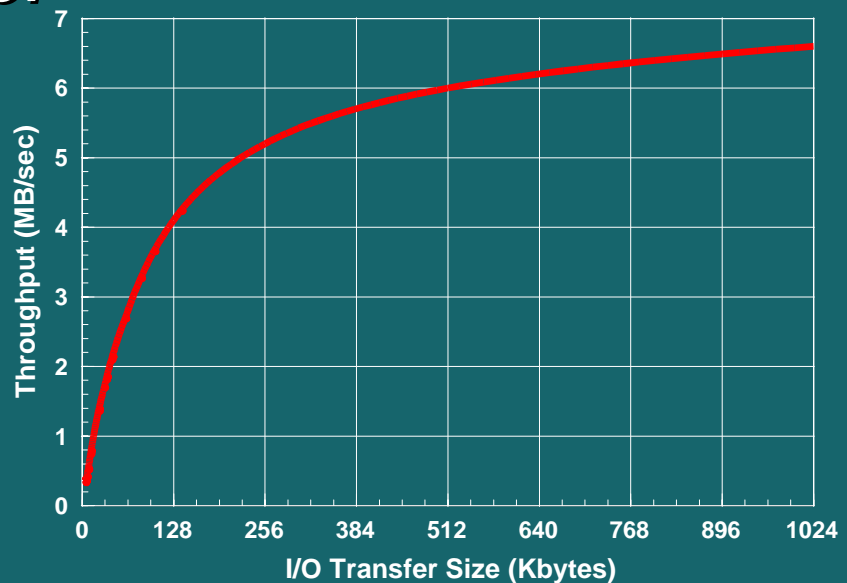
- *File system node assigns deadlines to real-time I/O based on stream data rate.*
- *Disk I/O is ordered by deadline.*
- *Quota of disk bandwidth reserved for non-realtime I/O.*
- *Non-realtime I/O is interleaved with real-time according to quota.*

Large Block Transfers

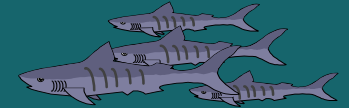


IBM Almaden Research Center

- *Tiger Shark uses large disk blocks (typically 256K, configurable) to increase throughput.*
- *Uniform block size within a file system makes I/O time predictable.*



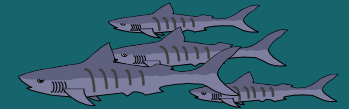
Load Balancing



IBM Almaden Research Center

- *Small number of “hot” titles may receive high percentage of system load*
 - *1000 MPEG2 streams = 800 MB/sec = 160 disks*
- *Alternatives:*
 - *replicate file on 160 disks*
 - *stripe files across 160 disks*
 - *cache in RAM*

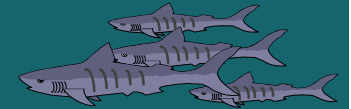
Load Balancing.....



IBM Almaden Research Center

- *RAM caching is cost effective only for very large servers*
 - *Crossover point varies with RAM cost and access skew, but usually is over 10,000 streams.*
- *Replication requires extra space for copies, extra bandwidth to make copies.*
 - *Load against each file varies with time.*
 - *Replication must be done continuously.*

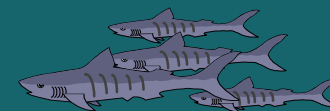
Wide Striping



IBM Almaden Research Center

- *Each Tiger Shark file system can be striped across many disks (on multiple storage nodes on SP-2).*
- *Wide striping provides inherently balanced load, efficient use of space.*
- *Tiger Shark fault-tolerance mechanisms ensure data integrity.*

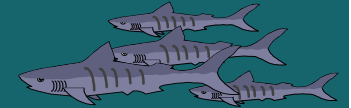
Disk and Storage Node Failure



IBM Almaden Research Center

- *Tiger Shark supports block-level replication or dual-ported RAID*
- *Block-level replication*
 - *All file system block pointers are n-dimensional arrays (n set at file system creation time)*
 - *Replicas allocated on separate disks and nodes*
 - *Any file can be replicated to any degree $\leq n$*
 - *Replication degree can be changed dynamically*
 - *Allows use of commodity disks (potentially cheaper than packaged RAID)*

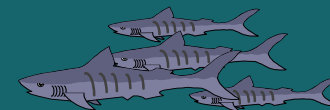
File System Node Failure



IBM Almaden Research Center

- *Metadata updates journalled to recovery log*
 - *Logs stored in special files on shared disk*
 - *Logs can be replicated for fault tolerance*
- *When a node fails*
 - *Surrogate node recovers from failed node's log*
 - *Surrogate releases failed node's locks*
 - *Recovery can proceed while other nodes run*
- *If log is lost, file system check utility is run to restore file system integrity*

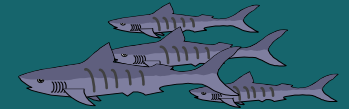
Read/Write Sharing on SP-2



IBM Almaden Research Center

- *Tiger Shark buffer manager provides cache coherency using Calypso token manager*
 - *byte-range tokens allow fine-grained sharing*
- *Posix read/write semantics - writes visible by other nodes immediately after write call completes.*

System Management

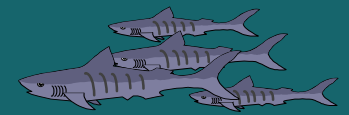


IBM Almaden Research Center

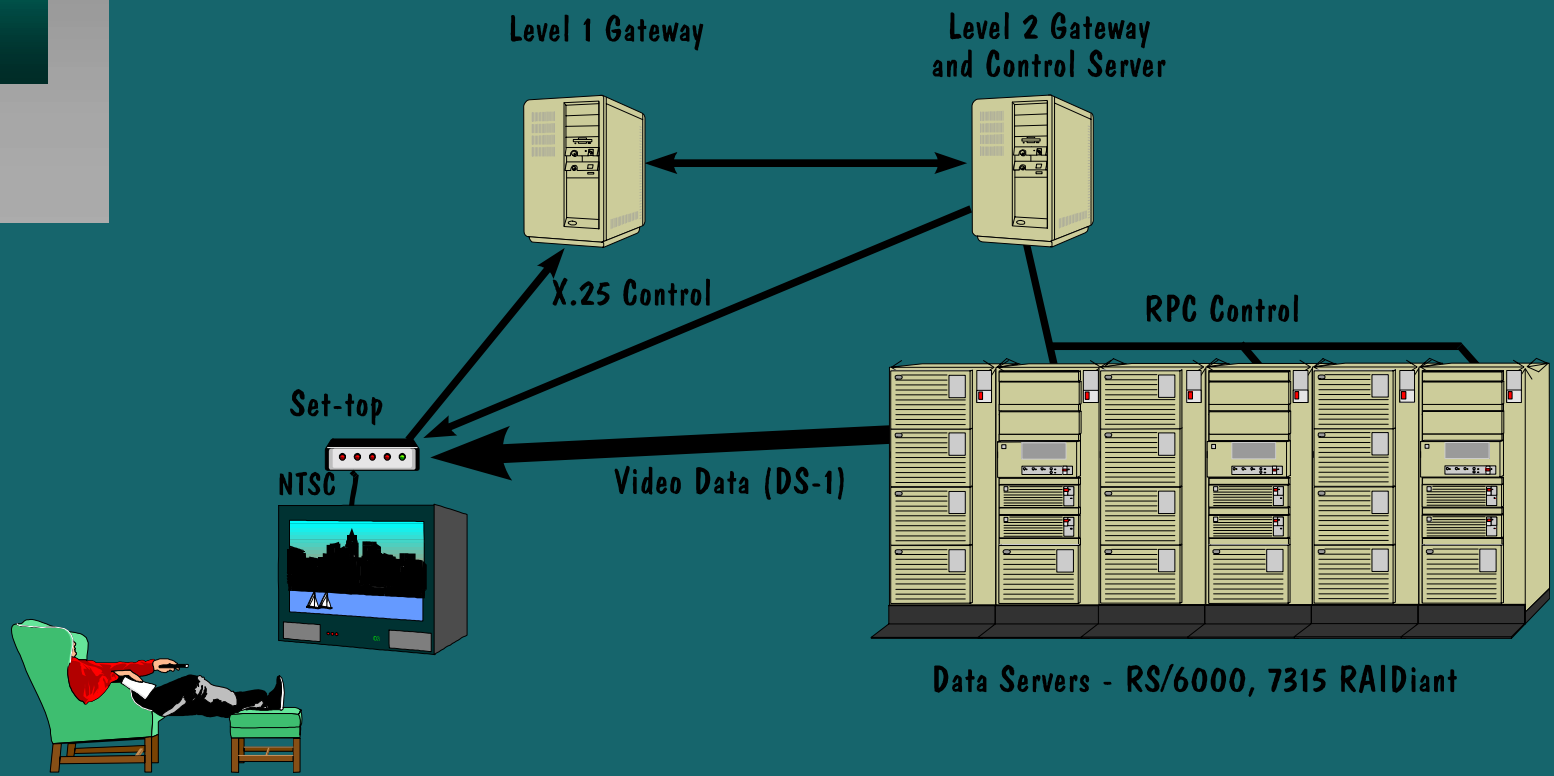
Online system management with no interruption of service:

- *Nodes can be added, removed, started, stopped*
- *File systems can be created or destroyed*
- *Disks can be added, removed, and replaced*
- *Files can be re-striped onto new disks*

Hong Kong Telecom VOD Trial

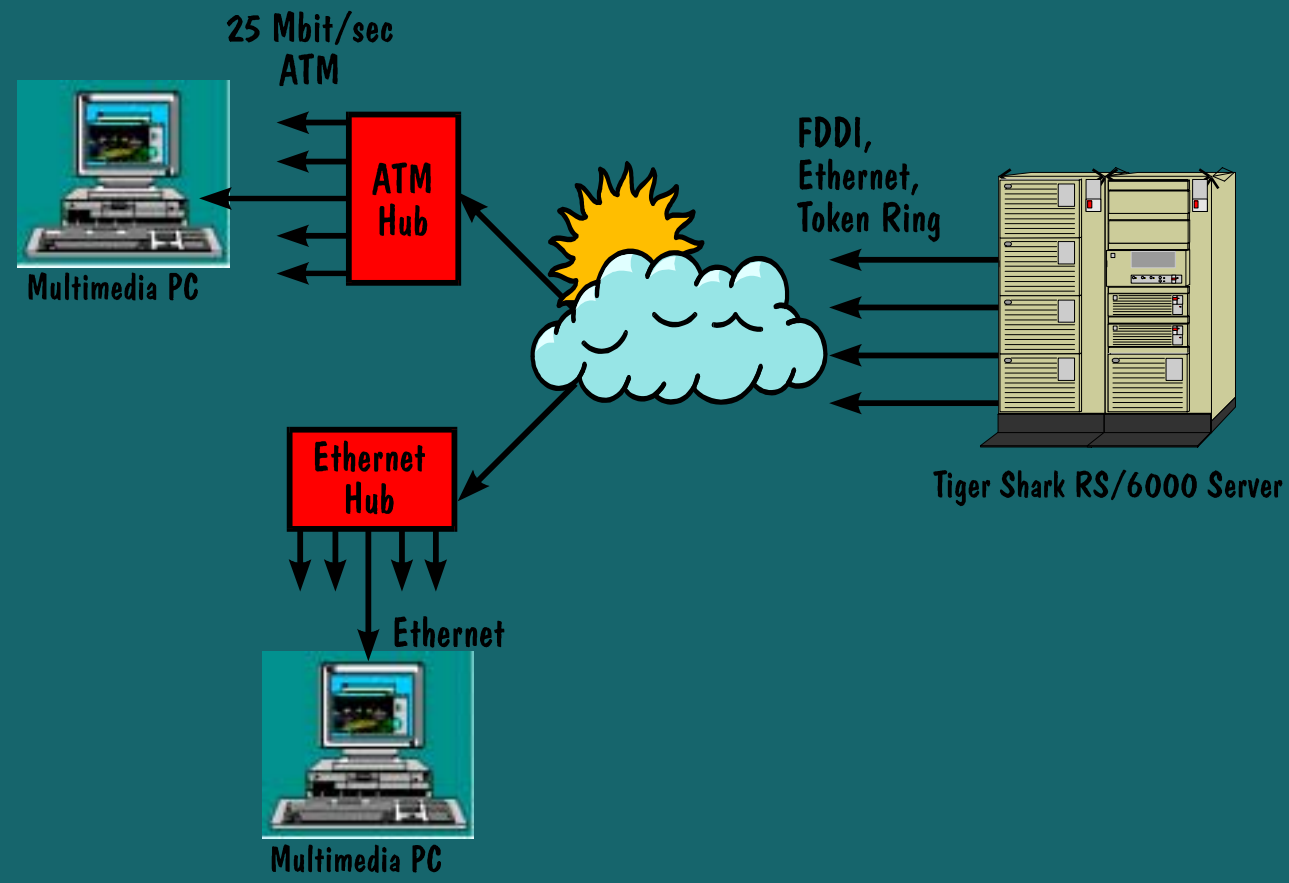


IBM Almaden Research Center



IBM Multimedia LAN Server for AIX

IBM Almaden Research Center

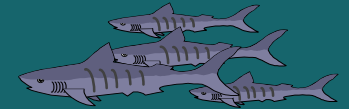


IBM Multimedia LAN Server for AIX

IBM Almaden Research Center

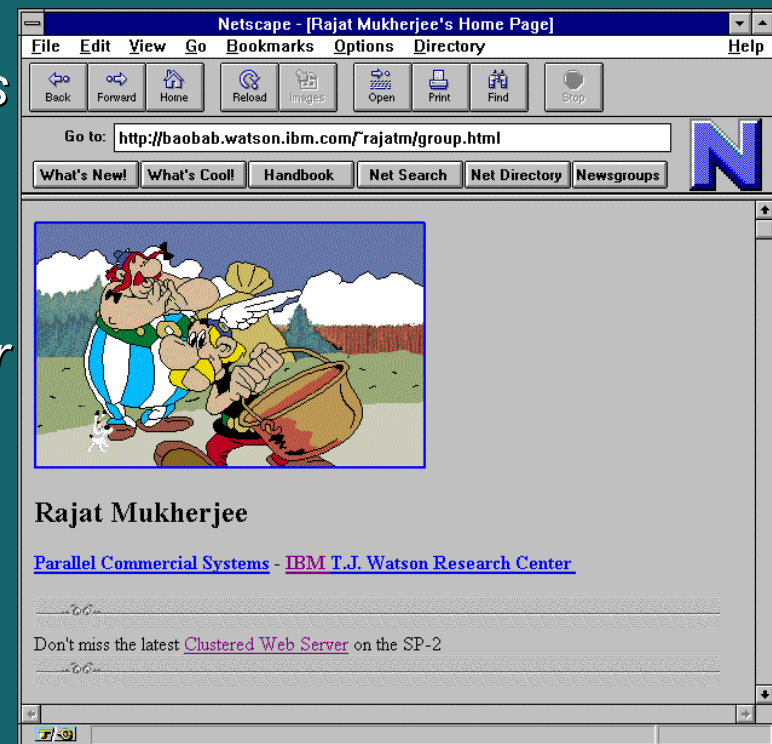
- *PC-based video applications use **file paradigm** to access video, audio on local CD-ROM or disk*
- *Multimedia LAN Server allows applications to access files on a server as if they were local*
- *Tiger Shark provides file system QOS*
- *Network QOS via high-performance LAN (e.g. switched Ethernet, ATM)*
- *File access via NFS protocol - supported by wide variety of clients (PC, Mac, workstations)*
- *Future protocol extensions as LAN, WAN support QOS*

SP-2 Scalable Web Server

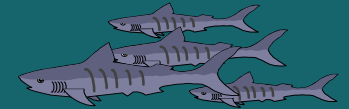


IBM Almaden Research Center

- *SP-2 based Web Server using Tiger Shark*
- *Applications: corporate servers (e.g. www.ibm.com), Web service bureaus*
- *Both conventional and multimedia data stored in Tiger Shark*
- *Streaming large files (e.g. software) similar to video*
- *Tiger Shark balances load against small files*
- *Fault-tolerance important for commercial Web applications*



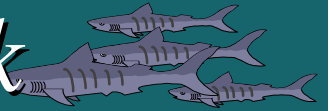
Summary



IBM Almaden Research Center

- *Tiger Shark is a scalable, open solution for video storage and playback*
- *Tiger Shark on SP-2 provides scalable platform for Interactive Television*
- *Tiger Shark/NFS on RS/6000 or SP-2 provides a scalable, high-performance platform for networked multimedia*
- *Tiger Shark can be used for non-multimedia applications (parallel computing, WWW)*

More Information on Tiger Shark



IBM Almaden Research Center

- *Tiger Shark Home Page*
 - *<http://www.almaden.ibm.com/cs/shark/>*
- *IBM Multimedia LAN Server for AIX*
 - *<http://www.austin.ibm.com/Cover/Announce.960220/>*
- *Internet Video*
 - *<http://www.rs6000.ibm.com/Cover/super95/sc95d.html>*
 - *<http://lscftp.kgn.ibm.com/pps/vibm/show/super95/demos/video.html>*