Scalable Network Storage at NASA

John Lekashman

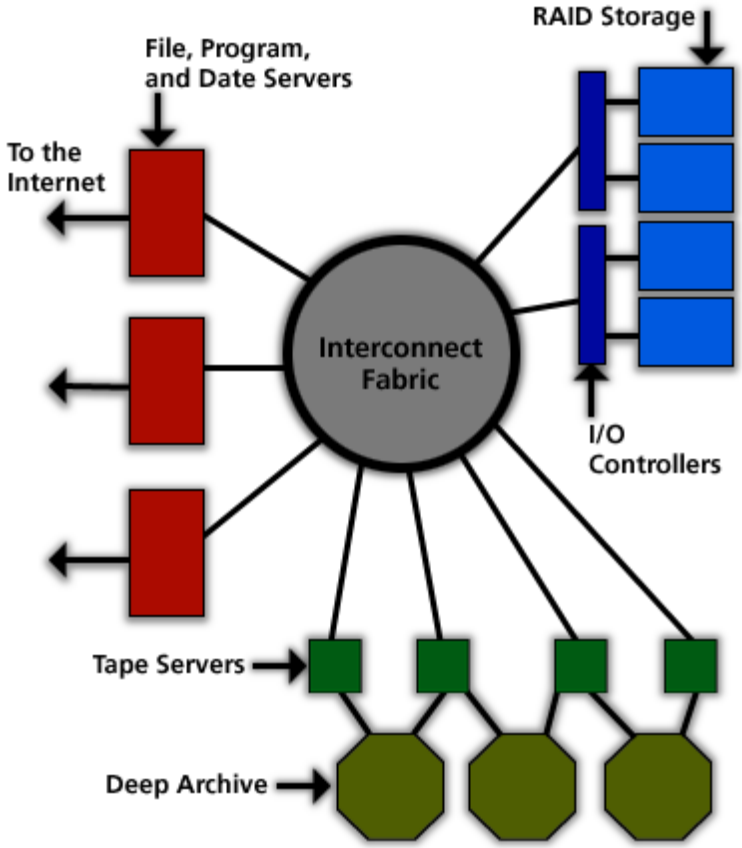
lekash@nas.nasa.gov

3/20/98

Storage - Summary

- **Objective**
  - ➔ To provide low cost, high performance, scalable data storage for long term data archiving, short term visualizations of very large data sets, and emerging data mining and analysis requirements.
- **Approach**
  - ➔ Develop new hardware based shared file systems in cooperation with University and commercial sector researchers. Use commodity components running a free, platform independent UNIX operating system (HetBSD, Linux). Develop fully distributed platforms on which to run such shared file systems. Expand the capability of commodity PCs by integrating new technologies (HIPPI, Fibre Channel, etc.) Create common software interfaces that will permit the use of interchangable hardware and software components.
- **Impact**
  - ➔ Commodity based systems can be readily scaled to meet both small and large storage requirements at optimal costs. The design allows storage rates in excess of 10 terabytes/day, a factor of 10 faster than current systems, at a factor of 8 less in cost.

The Generic Picture



Mission

- **Design and provide distributed systems that work at very high performance .**
  - ➔ **Provide technically advanced systems.**
  - ➔ **Perform the key integration and R & D steps we need so they can be built.**
  - ➔ **Make sure we understand the problem well enough so the systems scale.**
  - ➔ **Reference current size of the problem space:**
    - **Average current new data: 100 - 200 GB/day.**
    - **Current data store : 106 TB.**
    - **File count: 10 M.**
    - **Doubling period: 11 months.**

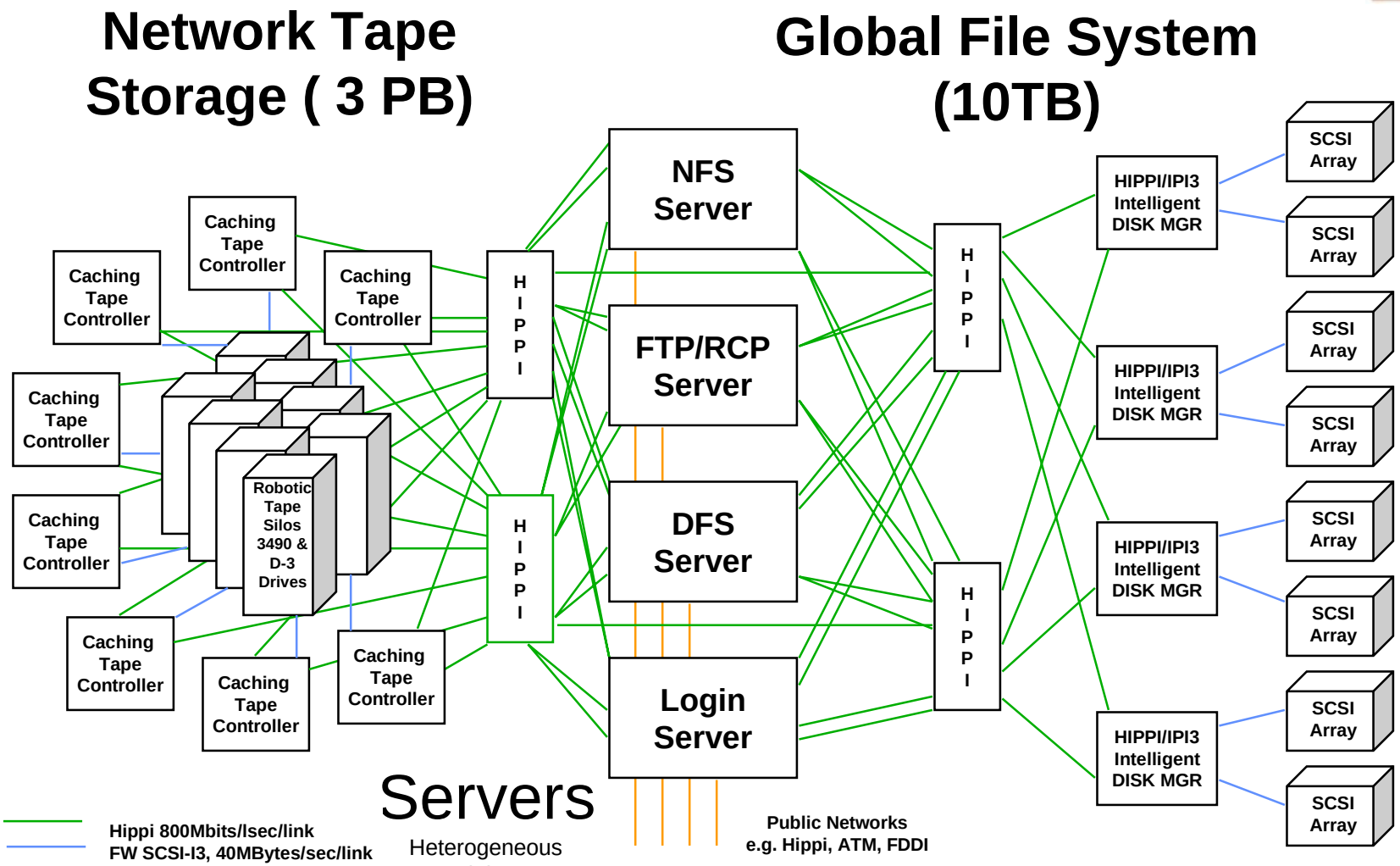
Strategy

- **Target R & D for systems that will handle that load.**
  - ➔ **Build networked mass storage systems, from commodity parts, because cost matters.**
  - ➔ **Work on data handling models for systems of this size.**
  - ➔ **Create network operating systems, to provide a reference implementation base.**
- **The key is system scalability.**
  - ➔ **What we create needs to scale, (at least) two orders of magnitude both up and down of the current target systems.**

Strategy II

- **Develop standard APIs for network storage.**
  - ➔ **Data Migration Interface Group (DMIG) - for creation of portable, interoperable data movers. (e.g. Hierarchical Storage Management).**
  - ➔ **GFS - for network disk file/block locking.**
  - ➔ **Network Tape handling, maybe Openvault.**
  - ➔ **Of course, heavily leverage those that work, Fibre Channel, IPI-3, TCP/IP, NFS v3.**
- **Control of the API defines who owns your data. You, or someone else.**

NAStore 3.0 System Design



Storage Systems

- **Create reference software implementation of distributed, multi-vendor, commodity mass storage.**
  - ➔ **NAStore - provides an application to validate feasibility.**
    - **Global File System, for fault tolerance and high performance.**
    - **DMIG Hierarchical Storage Management application.**
    - **NetBSD, for an open operating system implementation.**
    - **SGI IRIX, to show it with a vendor operating system.**
    - **Distributed metrics gathering.**
    - **Database development.**
    - **Virtual tape volumes for tape system buffer management.**
    - **Distributed volume manager for shared network tape robots.**
  - ➔ **Place into service, handling data storage for the facility.**

Related Research Activities

- **Operating systems R&D**
  - ➔ **NAS supports the "Free Operating System" research efforts.**
    - **NetBSD,FreeBSD, BSDI, OpenBSD, Linux.**
  - ➔ **NetBSD is our principal development platfor. It provides a portable, distributed operating system in which software R&D can take place.**
    - **Multi-threaded kernel.**
    - **Creation of a public, portable SMP operating system.**
    - **Development of portable, high performance device drivers.**
      - ⇨ **Target performance : 100K I/O transactions/second, without locking up.**

External Working Relations

- **Universities: (Current, active, funded)**
  - ➔ **University of Maryland BC - Large scale storage system simulation.**
    - **Working with Ethan Miller using trace data from both NASA Ames and NASA Goddard data sets.**
  - ➔ **University of Minnesota - Global File System.**
    - **Working with Matt O'Keefe on this, targeting (at least) both SGI and NetBSD interoperable implementations.**
  - ➔ **MIT - NetBSD development.**
    - **Working with Charles Hannum on public, portable multi-threaded SMP operating systems.**

Reference materials

- **We use the web for most of this.**
  - ➔ **It starts at: http://science.nas.nasa.gov/Groups/NAStore**