# "New Architectural Paradigms for Multi-PetaByte Distributed Storage Systems"

**Richard R. Lee**
Data Storage Technologies, Inc.
Post Office Box 1293
Ridgewood, New Jersey USA 07451-1293
Tel: +1-201-670-6620
FAX:+1-201-670-7814
e-mail: rrl@dst.com

**Abstract:**

*In the not too distant future, programs such as NASA's Earth Observing System, NSF/ARPA/NASA's Digital Libraries Initiative and the Intelligence Community's (NSA, CIA, NRO, etc.) mass storage system upgrades will all require multi-petabyte (or larger) distributed storage solutions. None of these requirements, as currently defined, will meet their objectives utilizing either today's' architectural paradigms or storage solutions. Radically new approaches will be required to not only store and manage these veritable "mountain ranges of data", but to make the cost of ownership affordable, much less practical in today's (and certainly the future's) austere budget environment!*

*Within this paper we will explore new architectural paradigms and project systems performance benefits and $/PB of information stored. We will discuss essential "top down" approaches to achieving an overall systems level performance capability sufficient to meet the challenges of these major programs.*

## Foreword

Today's data center is growing at a rate of per year of 40% CAGR, without even factoring in the impact of new multi-media and imagery-on-demand applications. This means that someone with a 10 TB problem today will have a 100 TB problem in 2-3 years and a multi-Petabyte problem in 7-10 years. Many of the large data centers found today have multi-PetaByte problems already. Based on this growth new exponential factors must be defined in order to understand the magnitude of the problem. Based on new exponent prefixes defined in the past two years, we have compiled a listing for reference throughout our discussions.

**TeraByte:** $10^{12}$ Bytes of bitfile data
**PetaByte:** $10^{15}$ Bytes of bitfile data
**ExaByte:** $10^{18}$ Bytes of bitfile data
**ZettaByte:** $10^{21}$ Bytes of bitfile data
**YottaByte:** $10^{24}$ Bytes of bitfile data

**Near-term Programs with Storage Requirements in Excess of 1 PetaByte:**

Within the federal end-user community today there are a number of requirements for multi-PetaByte archival systems already. A number of these will be based on years
and years of data gathering by numerous earth resources and imagery satellites producing warehouses of bitfiles which will be made available to thousands of researchers worldwide. For purposes of our discussion we will profile a sampling of the more visible ones.

**NASA EOSDIS:** Part of NASA's "Mission to Planet Earth", EOSDIS is a 13 site (8 directly associated with the program and 5 affiliated) distributed archive and data center for earth science data. This program has a data ingest, product generation & data distribution rate in excess of 1000 GB per day, with a 15+ year life span i.e. 11 PetaBytes anticipated over the program's life.

**NASA EDOS:** All incoming Level 0 data from EOS and International Partner satellite platforms is collected at this site in WV for archiving and processing into higher order data products. It is then distributed to the 13 EOSDIS sites (as well as IP sites upon request). Level 0 and higher order data products in excess of 1 TB per day will be archived, processed and distributed from this site over the 15+ life span of the program. Total archive capacity will exceed 1 PB during this time.

**NSF/ARPA/NASA Digital Libraries Initiative:** Envisioned as the "Data Malls on the Information Superhighway" these distributed information infrastructure servers will provide fast access to thousands of TB's of data, and will open the infrastructure up to the general public. They are intended to capture, store, distribute and provide access to every type of bitfile data available from public and private sources. Given the scope of this plan it is envisioned that this will comprise hundreds to thousands of PB's over its useful life.

**NII - "High Resolution Video on Demand Services"** As one of the most visible components of the National Information Infrastructure concept, this application has been embraced by the entire telecommunications and computer industries as well as capturing a significant share of the NII federal funding dollars available, and the public's mindshare as well.

Using the most advanced image compression techniques available today can only reduce the large size of a "digital movie" to 10's of GB's (assuming higher resolutions than found in conventional broadcast today). This nets out to a requirement of multiple PB's in key VOD locations serving major metropolitan areas across the country (each Blockbuster Video location currently houses in excess of 10,000 feature length movie titles).

**The Intelligence Community's Consolidation of Disparate Archives:** Hard to describe in any other terms, the United States' Intelligence Community (CIA, NSA, NRO, DIA, etc.) is faced with dilemma of providing higher and faster levels of service to its end-users with less capital to work with (dollars and personnel). In total, the IC ingests over 4 TB per day from classified sources alone (1 TB+ per day from images that are approximately 1 GB each), not to mention the thousands of unclassified sources worldwide that are routinely accessed. In trying to meet the needs of their end-users they must respond to numerous real-time queries across disparate resources. All combined, the IC has in excess of 10 PB of data already archived, with this growing at a much higher rate than that of the rest of end-user community (60+% CAGR).

**Current Storage System Architectural Paradigms i.e. "Multi-TeraByte Class":**

**Types:**

- Direct Connected Peripherals i.e. "The Mainframe Era"
- Stand-alone Data Servers i.e. "The Client/Server Mantra
- Network Attached Peripherals i.e. "The NSL Approach"
- IEEE Mass Storage Systems Reference Model i.e. "The Open Systems Standard"

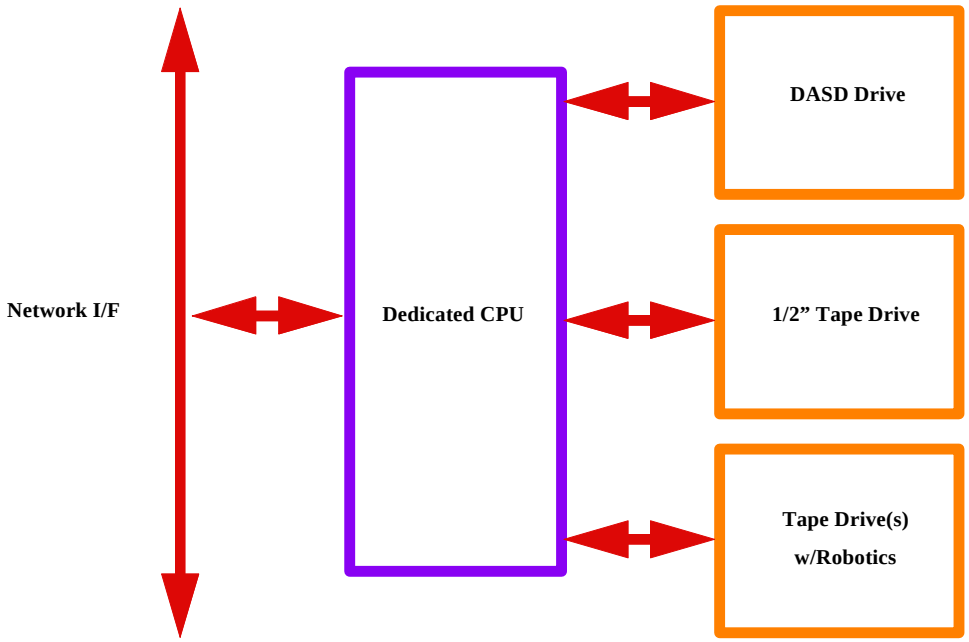


**Figure 1: “TeraByte Class” Data Server**

**• Overview:**

Systems conforming to the first two of these types of architectural paradigms (Mainframe attached and Client/Server) are essentially CPU Centric and appear as a centralized repository of bitfiles to the outside world. Bitfiles may be distributed out over a network to various clients, but still originate from a centralized location. The IEEE Mass Storage Reference Model promises to break this scheme into either a distributed or quasi-distributed one, but all implementations fielded to date behave in a centralized manner and will potentially fall apart when distributed.

In short, these systems all suffer from the same type of performance limitation; that of acting as a single point of access for all classes of service. The controlling/serving CPU can only maintain one connection/DMA access at a time in practical terms and even through the use of multiple CPU’s and multi-threaded OS’s one can only maintain a small number of transfers simultaneously (mostly due to shared memory and operating system software limitations) appearing on an effective basis as a single point of access to the network.

Systems based on the NSL/HPSS paradigm (network attached peripherals) are designed to support high-speed transfers of large bitfiles, but do not translate to a distributed environment and are far too costly for the mainstream of the end-user community. For this reason we have classified them as part of the TeraByte class.

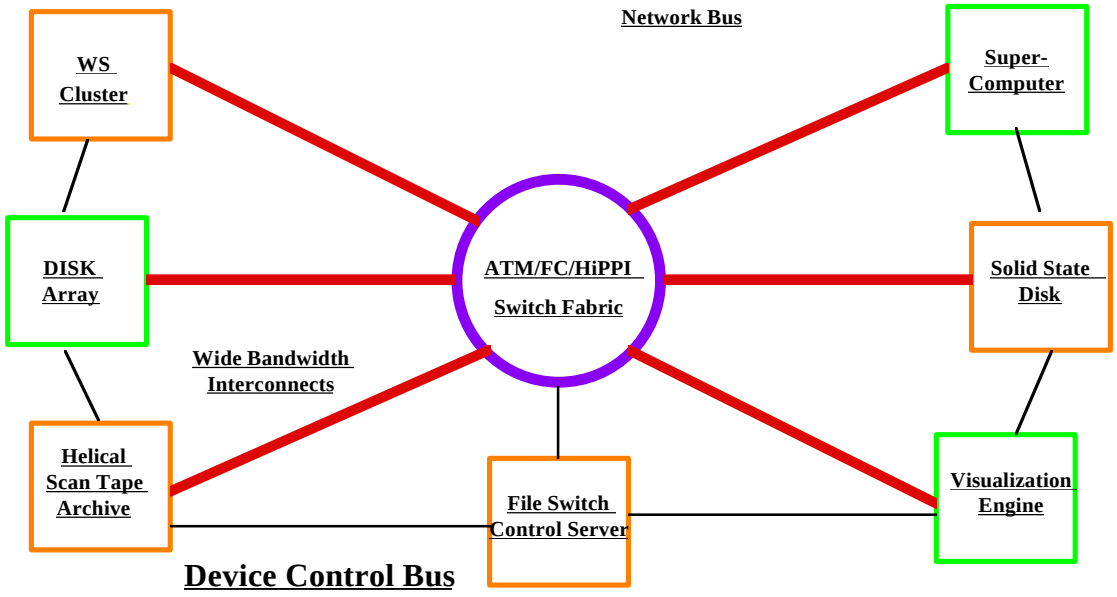


**Figure 2: "The NSL/HPSS Paradigm"**

## Concepts for Future Storage System Architectural Paradigms: "Multi-PetaByte and Beyond"

In order to meet the challenges of managing multi-PetaByte distributed archives we need to think beyond the current COTS mindset and explore new approaches altogether, some based on concepts being used in parallel computing today (using however, COTS components where practical). We feel that a parallel architecture eliminates much of the problem encountered with "single point of access" found in traditional architectures of the day. Much of what we will present is still in the early stages of development, but does represent a logical approach to the problem at hand.

**• Distributed Cluster-type:**

This architecture envisions an environment where a clustered array of servers are interconnected via a LAN to a series of data repositories. These servers are in turn connected to a WAN and serve clients and other servers distributed throughout the enterprise. Each repository contains multiple peripherals and robotics assemblies for contention free search and access of bitfiles. Using fast packet technology, the system is capable of storing and retrieving bitfiles within the repositories at very high packet rates, but at a relatively low cost. Utilizing this type of architecture allows for many points of access, while retaining the benefits of using commodity type technologies.

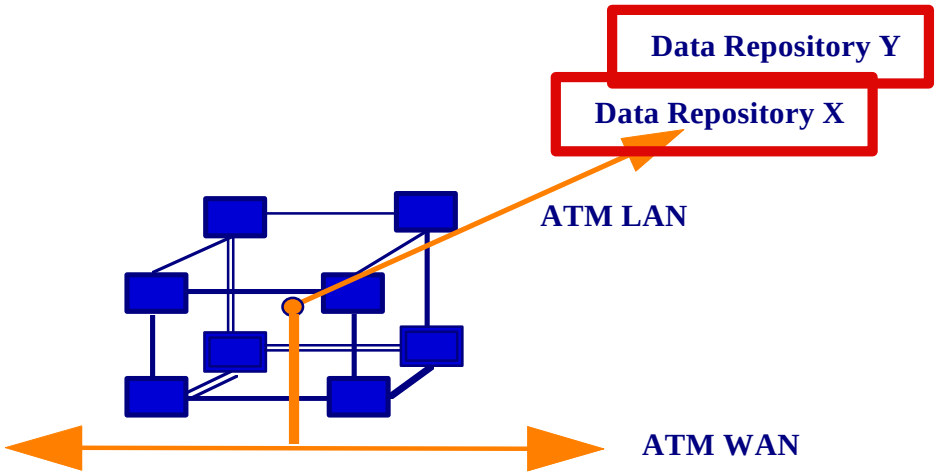


**Figure 3: - A Distributed "Cluster -type" Query/Bitfile Server**

**• Scalable Parallel**

This architectural approach borrows much from today's scalable processors i.e. shared memory parallelism. The system is essentially demand driven and each process automatically adapts itself to the number of resources (CPU's and peripherals) available to the user at the time of the request. This architectural approach is totally scalable and higher levels of performance can be obtained by merely adding more CPU's and peripherals i.e. forward extensibility without obsolescence.

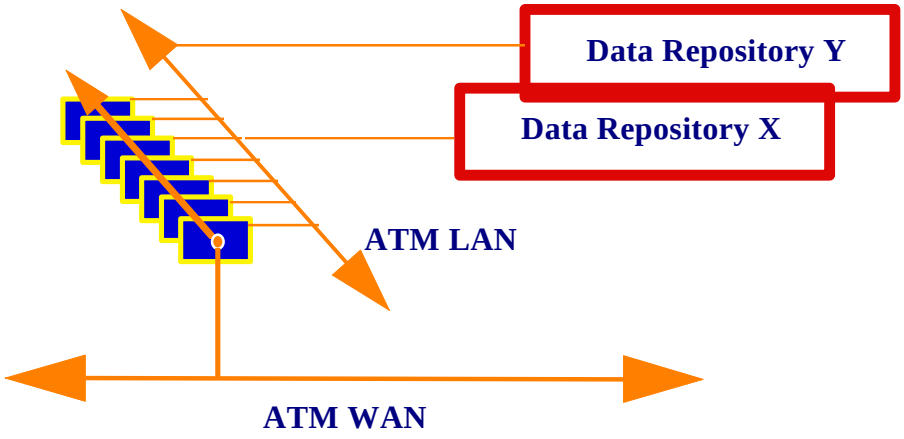


**Figure 4: - "A Scalable Parallel Query/Bitfile Server"**

**• Dynamically Configurable**

As implied by its name, this architectural approach is the most flexible in meeting "data on demand" requirements. The system configures itself dynamically depending upon end-user demand and resources available. During times of extremely high demand the system configures itself as highly parallel, while during periods of light-medium demand it acts as a clustered resource. The benefits of this approach are that it eliminates single-node bottlenecks (the slowest component of a distributed system throttles the performance of the entire system) and acts as a high-availability resource under all load conditions.

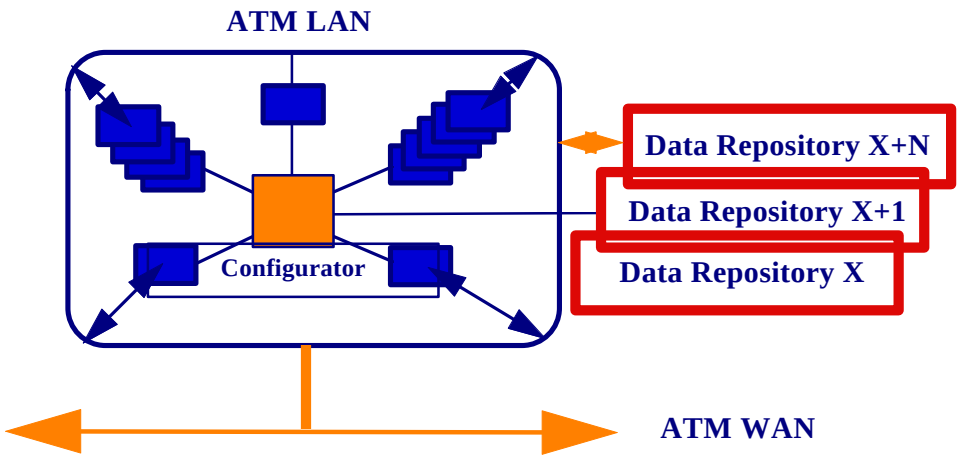


**Figure 5: - "A Dynamically Configurable Query/Bitfile Server"**

The concepts that we have discussed here are by no means new or all encompassing. Rather, they are shown as examples of wide departures from the status quo which seems to pervade the mindset of today's systems planners and developers as the only approach available to meet the challenges set forth. We expect that as everyone's eyes are opened wider to both the scope of the challenge as well as the tools available to respond to it, that new mindsets will develop.

**Additional Considerations:**

Adoption of new hardware architectural paradigms alone will not suffice to meet the challenges of these ever increasing requirements. We will need to accomplish the following in parallel with these developments;

• Adopt Object Driven files systems for faster query, search and access to bitfiles
• Continue to develop “bandwidth on demand” driven internetworks and storage peripherals
• Eliminate all “single point of access” failures and bottlenecks
• Utilize distributed Metadata and Browse data db’s
• Migration to higher order data transfer and communications protocols
• Achieve continuing incremental reductions in Unit Storage Costs with attendant increases in Capacity-per-physical unit and vastly improved data reliability.

**Cost Projections & Realities:**

Based on the use of conventional architectures and components, we project that most end-users are looking at fielded system costs of $40-60M per PetaByte, with the majority of these costs being centered around expensive CPU’s, network fabrics and high-end peripherals. This level of cost is far too high for most, if not all budgets today and does not include the manpower or materials necessary to operate and maintain these systems over their useful life (a major component of total cost).

We believe that in order for the key programs discussed earlier to be achievable, that costs in the $10-20$/PetaByte range must be achieved. This can only be realized by embracing radical new approaches similar to what we have outlined.

**Conclusions and Recommendations:**

Current architectural approaches “bottom out” when tasked at multi-PetaByte levels (access, bandwidth, file management, cost, etc.).

Scalable and dynamically Configurable hardware architectures off significant promise in overcoming many of these limitations.

In addition, exponential increases in hardware, software and protocol efficiencies are mandated to meet this challenge as well.

In short, “The ways of the past must give way to the needs of the future” i.e. the familiar and comfortable path of the present will not suffice.

**References;**

[1] Lee, R. and Dan Mintz, "Grand Challenges in Mass Storage - A Systems Integrators Perspective", Second NASA Goddard Conference on Mass Storage Systems and Technologies, Greenbelt, MD, September 1992

[2] Lee, R., "The Future of Mass Storage", THIC Winter Meeting, San Diego, CA, January 1993

[3] Lee, R., "New Architectural Paradigms for Multi-PetaByte Distributed Storage Systems", Massive Digital Data Systems Workshop, Reston, VA, February 1994/Supercomputing '94, Washington, D.C., November 1994

[4] Kuhn, T., *"The Structure of Scientific Revolution"*, University of Chicago Press, Chicago, IL 1970

[5] Lee, R., 19mm Helical Scan Recording Technology for Data Intensive Computing Environments", 10th IEEE Symposium on Mass Storage Systems (vendor poster session), Monterey, CA, May 1990

[6] Coleman, S. and R.W. Watson, "The Emerging Paradigm Shift in Storage System Architectures", review copy for Proceedings of the IEEE, April 1993

[7] Coyne, R. , H. Hulen and R. Watson, "Storage Systems for National Information Assets", Proceedings- Supercomputing '92, Minneapolis, MN, November 1992

[8] Lee, R., "Mass Storage - the key to success in high performance computing" , Convex File Server Seminars, Milan/Rome, Italy, February 1993/Third NASA Goddard Conference on Mass Storage Systems and Technologies, College Park, MD, October 1993

[9] Lee, R., "19mm Data Storage Applications", THIC Fall Meeting, Annapolis, MD, October 1990

[10] Panel Discussions, Mass Storage Roundtable, Supercomputing '94, Washington, D.C., November 1994

[11] EOSDIS Core System Science Information Architecture "White Paper" Doc #FB9401V2, Hughes AIS, Inc., Landover Maryland, March 1994

[12] Dixon, Dick, "Statement of Requirements of the European Mass-Storage Specification Working Group Working Version 1.1" , European Weather Centre, June, 1994

[13] Teaff, Danny, "The High Performance Storage System", IBM U.S. Federal Publication

[14] IEEE Storage Systems Standards Working Group, "Mass Storage Systems Reference Model Version 5", IEEE Computer Society Mass Storage Systems and Technology Committee, Balloting Draft, July 1994

[15] "National Science Foundation's MetaCenter", NSF Division of Advanced Scientific Computing, NSF Publications, Arlington, VA., 1994

[16] “Program Guideline/Program Briefing”, ARPA/NASA/NSF Research on Digital Libraries Initiative, Arlington, VA, September/December ‘93

[17] Convex Exemplar System Overview, DOC 080-002293-000 V1.1, Convex Computer Corporation, Richardson, Texas, 1994