IEEE Standards
Storage Systems

Jack Cole

Sponsor
Storage System Standards
Working Group

Army Research Laboratory
Aberdeen Proving Ground, MD

30.3.2000 Jack Cole/ARL

Jack Cole is both the IEEE Computer Society (CS) Standards Activity Board (SAB) Chair for the Storage System Standards Committee (SSSC), and Chair for the Storage System Standards Working Group (SSSWG). The SSSC Chair is the sponsor of SSSWG. Merritt Jones was this sponsor until recently. A new chair will be selected for SSSWG.

GOAL

Today

Inform About

IEEE Standards Efforts

Status of IEEE MMS Standards

Generally

Hasten Convergence
of Competing Technologies

Foster Transparent Information Access
In Distributed, Heterogeneous
Computing Environments

Sponsor Message

- **Army Research Lab (ARL)**
- **Aberdeen Proving Ground, MD**
  105 km Northeast of College Park
- **Formerly BRL … ENIAC … BUMP (DMF) … "ping" … Army's First Supercomputer**
- **DoD HPCMP MSRC**

30.3.2000 Jack Cole/ARL

"BUMP" Stands for "BRL-USNA Migration Project" on which the Cray Datamigration Facility (DMF) is based. "USNA" is the US Naval Academy

Mike Muuss/ARL is the author of the UNIX network command "ping".

The Army's first supercomputer was a CRAY-2 in the mid 1980s, followed shortly afterward by a CRAY X-MP. The presenter was the system administrator for these systems which were shutdown in the early 1990s.

Initial work to startup the DoD High Performance Computing Modernization Program (HPCMP) began with Tony Pressley and his 'technology integration team' (of which the presenter was a member) writing the statement of work (SOW) and other documents. Tony was the first Program Manager (PM), and now works at ARL in the HPC Division. The present HPCMP PM is John Baird, and the HPCMP has four major shared resource centers (MSRCs) and perhaps a hundred distributed centers (DCs), with some DCs being quite large (e.g., the Maui HPC Center (MHPCC)).

ARL http://www.arl.mil HPCMP http://www.hpcmo.hpc.mil

IEEE Computer Society

- Technical Activities Board (TAB)
  Mass Storage Systems
  Technical Committee
  http://computer.org/tab/tcms/

- Standards Activity Board (SAB)
  Storage Systems
  Standards Committee (SSSC)
    Storage System Standards
    Working Group (SSSWG)

30.3.2000 Jack Cole/ARL

Background

- **SSSWG Beginnings, Charter**
- **Guide for Storage System Design, Mass Storage System Reference Model (MSSRM)**
- **Change of Direction from MSSRM**
- **2000: Tenth Anniversary of SSSWG**

30.3.2000 Jack Cole/ARL

The Storage System Standards Working Group (SSSWG) Charter is dated April 3, 1990, presented at the Third IEEE MSSTC Storage System Standards Workshop held in Monterey, CA in conjunction with the Tenth IEEE Mass Storage Symposium on May 10, 1990. "**MSSRM**" is the "Mass Storage System Reference Model" version 5 was last revised and approved internally by the SSSWG in September 1994. The components (SOID, PVL, PVR, MVR, MGT, VSS) are described in the MSSRM which may be view at the SSSWG web site http://www.ssswg.org. **Excerpts from the Charter**: The IEEE Storage System Standards Committee is chartered to model generic mass storage systems, and based on such modeling, to develop widely accepted, readily implemented standards with minimal licensing requirements. In addition to working on standards, the SSSWG may develop recommended practices and guides. The SSSWG is primarily concerned with Distributed Storage System Design, and the SSSWG, without favor, includes Storage Systems of every scale in its studies. An object-oriented approach is desired in all SSSWG efforts, and net-attached storage is intrinsic to its model. SSSWG must also consider promising emerging technologies in its modeling, even though standards for parts of the model may not be immediately practical as a result.. The purpose of these standards is to promote use of best technologies resulting in interoperable, fully-scalable systems permitting ready access of information throughout distributed, secure, heterogeneous net-attached storage systems.

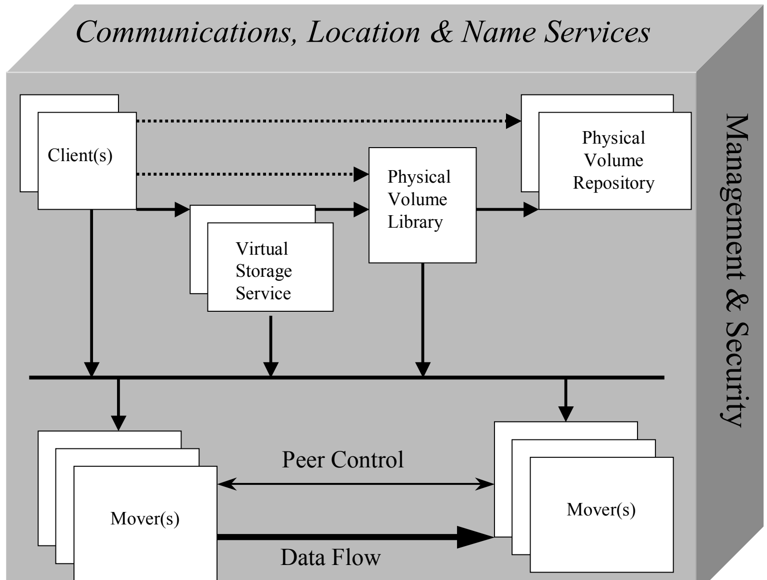
"Guide" and MSSRM

- The "Guide" Updated, Re-Approved IEEE Project this Year
- 1994 MSSRM is valuable Un-Balloted "Guide" viewable at *http://www.ssswg.org*
- Components - PVL, PVR, MVR, VSS

30.3.2000 Jack Cole/ARL

"Guide" PAR This project will produce a clear, abstract, model exposing the design features required for storage systems to provide transparent, secure information access in highly distributed, heterogeneous computing environments. The model produced will describe design alternatives and rationales applicable within the spectrum of valid storage system architectures. Emphasis in the model will be placed on net-attached storage, object-oriented design, open source software, minimal licensing alternatives, and maximum scalability. The work under this project will serve to revise the popular IEEE Mass Storage System Reference Model version 5 of 1994, and use it as a basis for related IEEE Recommended Practices and Standards. Information is being generated at rates in step with the rapidly improving technologies of processors, and while storage media and hardware are improving rapidly as well, storage system software and architecture lag dangerously behind. At the same time demands for access to information in highly distributed and heterogeneous environments are growing. Traditional approaches are not sufficient to meet these needs, and revolutionary or rapid evolutionary changes are needed in storage system design and software. This Model will guide implementers toward interoperability in meeting such demands by suggesting best use of current and emerging technologies. This Model will inspire commercial designs of storage systems and system components from a broad spectrum of implementers, resulting in a high level of interoperability throughout the world.

The MSSRM



30.3.2000

Jack Cole/ARL

Standardizing PVL, PVR

- **MSSRM**
  Wealth of Ideas
  Basis of Present Systems
  Retains Active Interest
- **SGI OpenVault ™ Base for IEEE Media Management System (MMS)**
- **Update MSSRM based on MMS**

30.3.2000 Jack Cole/ARL

The MSSRM pre-saged net-attached storage and separation of control and data flow, features of storage systems today. The Physical Volume {Library | Repository}, Data Mover components are used as models of commercial products today. Active interest in the MSSRM continues at the SSSWG web site.

SGI OpenVault™ developers attended a SSSWG meeting in November 1996 which Dixon Hutchinson arranged, and as a result, SGI presented OpenVault™ to the SSSWG in January 1997. A motion was made, and in one of the few official internal votes of SSSWG, the group agreed to the idea of beginning MMS standards based on OpenVault™. The group included the original PVL, PVR author, Rich Wrenn.

OpenVault™ took a pragmatic approach to the MSSRM. Where the MSSRM attempted to throw a net over all possible needs of a storage system, OpenVault™ only attempted to implement essential services, a minimalist approach. And OpenVault™ had been influenced by MSSRM ideas in PVL, PVR, as well as the OpenVault™ team containing two charter members of SSSWG (Mike Hardy and Lloelyn Cassell),

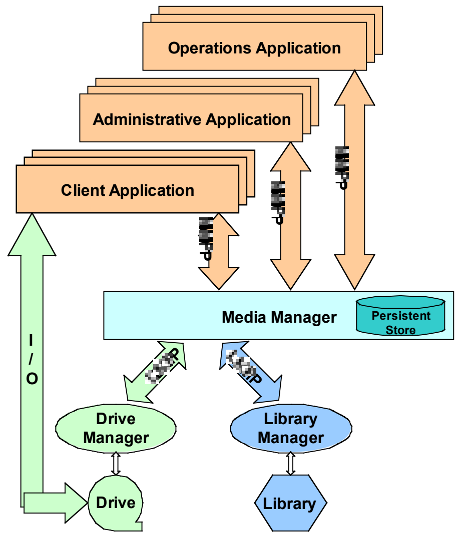
MMS Described

Tutorial March 1999, April 2001?

- MMS Characteristics
- MMS Suite of Standards
- Distributability of MMS

30.3.2000 Jack Cole/ARL

MMS Components



30.3.2000 Jack Cole/ARL

MMS Characteristics (1)

- Based on
  Open Source Implementation
- Defines Basic Functionality of MM
- Protocol, not API Based
- Neutral WRT
  - Media Type and Content
  - Platform and OS
  - Language of Implementation

30.3.2000 Jack Cole/ARL

MMS Characteristics (2)

- Fully Scalable, Modular
- Distributed
- Secure
- Application Independent
- Multiple Vendor Implementations Interoperable

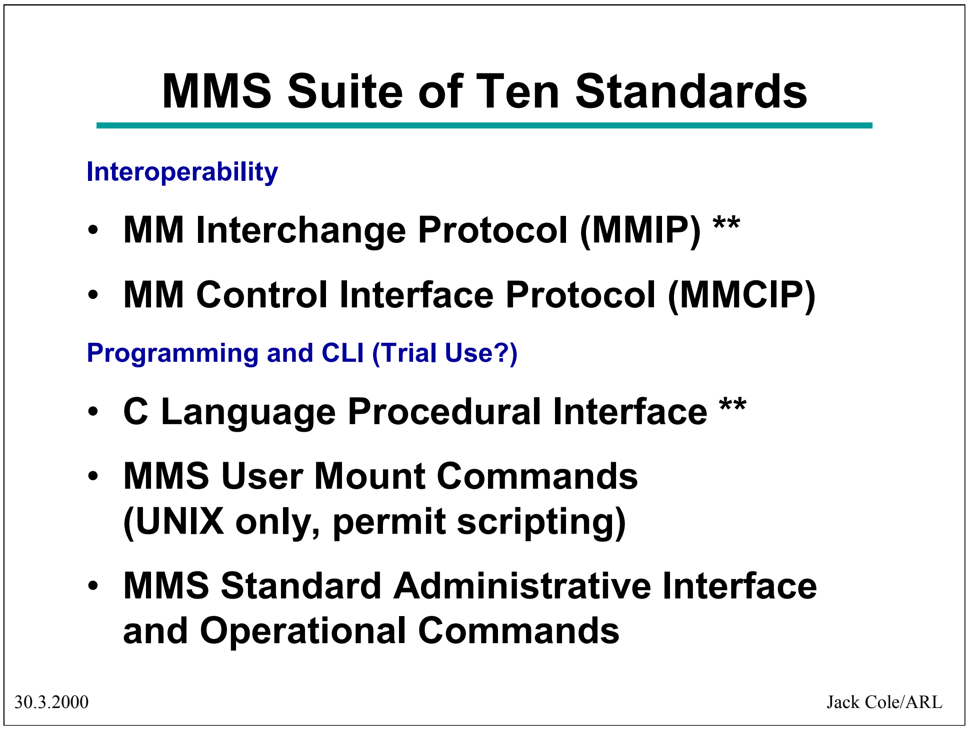
MMS Suite of Ten Standards

- **Architecture (and Data Model)**
- **Session Security, Authentication, Initialization Protocol (SSAIP)**
- **Media Management Protocol**
- **Drive Management Protocol (DMP)**
- **Library Management Protocol (LMP)**

First Five Balloted

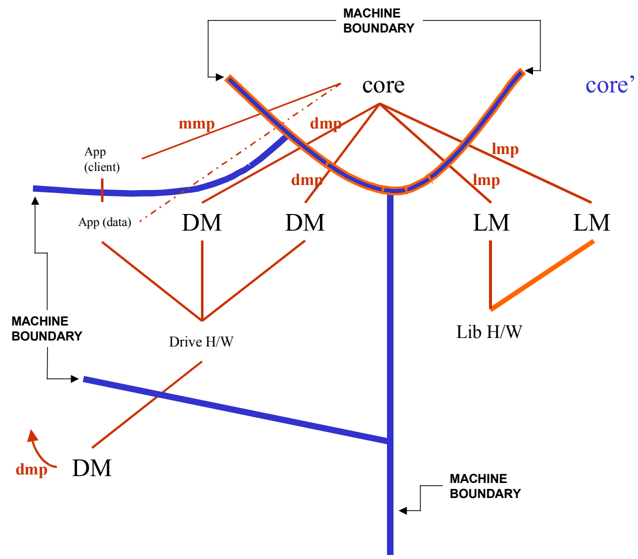
30.3.2000 Jack Cole/ARL

1244.1 - MMS Architecture. Specifies the architecture of a distributed, platform-independent, system to manage re-movable media, including both disk and tape, using robotic and manual methods. The general schema for managing media, the expected components of the software system, and the data model to be supported by the software system for managing this media are described by this standard. Details of components of the Media Management System are specified by companion standards. **1244.2 - SSAIP** is the initial "handshake" protocol used by components of the MMS to establish identity, authority, and initial communication. **1244.3 - MMP** used by client and administrative applications to allocate, deallocate, mount, and dismount volumes, and to administer the system. The MMP includes levels of privilege so that, for example, a client application cannot perform administrative functions, or an operator console program cannot perform higher-level management functions. **1244.4 - DMP** is used between two software components of the MMS: the central management core and a program that manages a drive which is used to access removable media. **1244.5 – LMP** is used between two software components of the MMS: the central management core and a program that manages an automated library or a vault. The minimum functionality required to implement an MMS is the SSAIP and MMP. Most practical implementations will include the DMP and LMP. Additional protocols are defined to extend the MMS to interoperate with other MMSes and with other media management systems:



1244.6 - Media Manager Interchange Protocol (MMIP) defines protocol to allow interchange of information between autonomous Media Managers. **1244.7 - Media Manager Control Interface Protocol (MMCIP)** defines protocol which permits interfacing the data management component of the MMS with existing library management systems. **1244.8 - C Language Procedural Interface** defines a set of standard programming interfaces which facilitate construction of components of the MMS, particularly client, administrative, and operational applications, library managers, and drive managers. The initial definition will be for the C programming language. The interface will be designed so that implementation in languages such as C++ or Java could be easily accomplished. **1244.9 - MMS User Mount Commands** defines a set of standard commands to allow user to mount, unmount, acquire, and release media. These commands are specified as a part of a command line interface for systems that offer such interfaces, such as the UNIX shell or NT command line interface. Commands may be embedded in scripts to produce more complex or custom functions, or to allow an application program that is not written for MMS to be adapted for use with MMS. **1244.10 - MMS Standard Administrative and Operational Commands** defines a set of standard administration and operation commands of an MMS. The standard defines a command- line, minimally interactive interface for basic interaction with the MMS; these commands could be used to construct interactive interfaces using scripting-based systems such as web CGI scripting or tcl/tk.

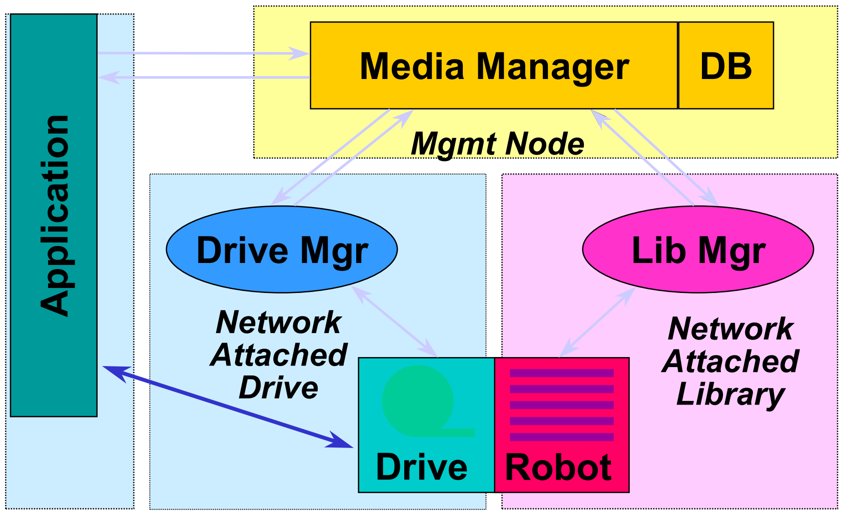
Distribution of MMS Components Across Machine Boundaries



30.3.2000

Cole/ARL

MMS in a Storage Area Network



30.3.2000 Courtesy of Geoff Peck, March 1999 Tutorial Jack Cole/ARL

Balloting First Five MMS Standards

- Balloting Took Place
  Dec 15, 1999 – Jan 14, 2000
- Total of ~ 300 Pages
- Received ~ 200 Comments,
  Predominantly Editorial
  from 36 "Balloters"

Balloters ...Thanks!

- **Balloting Pool ~53 , Group ~36**
- **Balloters Joined IEEE SA**
  some as out of pocket expense
- **Balloter Effort Significant**
  over holidays and through Y2K
- **Sensitivity of Balloters**
- **Internal IEEE "Coordination"**

30.3.2000 Jack Cole/ARL

The "Pool" consists of individuals who registered interest in receiving invitations to ballot. The "Group" is composed of individuals who asked for invitations and responded affirmatively to them.

Balloters, such as those working for the US Government, pay for IEEE membership themselves. Some of these people did so just to ballot these standards. And all of the balloters were forced to review and ballot through the Holidays and Y2K. Our gratitude for their work is quite high.

The group was "sent" official IEEE Ballots for the 1244 standards, with a month to ballot and submit comments.

No paper ballots were used. The draft standards were placed online in PDF format at http://www.ssswg.org, as well as copies being sent to IEEE. The IEEE Standards Department Balloting Service then sent e-mail to all members of the Balloting Group informing them of the SSSWG.ORG location of files and the IEEE.ORG location for balloting.

The IEEE Balloting Service received and tallied the ballots. And subsequently sent the SSSWG Sponsor the results and comments from Balloters and the IEEE Coordination groups. Among the latter is the IEEE Registry Authority Committee (RAC), which manages the IEEE OUIs (Organizationally Unique Identifiers).

Balloting Results

Draft Standard	No. Of Ballots	Balloting Results		
		Affirm	Neg	Abstain
P1244.1/Architecture	36	92%	8%	3%
P1244.2/SSAIP	35	91%	9%	7%
P1244.3/MMP	35	91%	9%	7%
P1244.4/DMP	35	100%	0%	16%
P1244.5/LMP	35	100%	0%	16%

30.3.2000 Jack Cole/ARL

The percentages shown for Affirm/Neg do not include abstentions. Abstention votes are considered valid for achieving the required ballot return rate of 75%, but are not considered when calculating the approval rate.

Kinds of Comments

- Editorial (predominant kind)
- Technical
- Technical Required
- Coordination Committee
- Many Easily Answered Regarding Objects, BNF, Oversights, …

30.3.2000 Jack Cole/ARL

Kinds of Responses

- **Adopted**
  majority response
- **Next Version**
- **Already Met**
- **No Plan to Address**

30.3.2000 Jack Cole/ARL

Sample Comments (1)

- **Data Model Needs Common Means of Counting Device Errors**
- **Could apply to Non-Removable Media**
- **Conformance Levels Needed**
- **Distributed Atomic Mounts Needed**

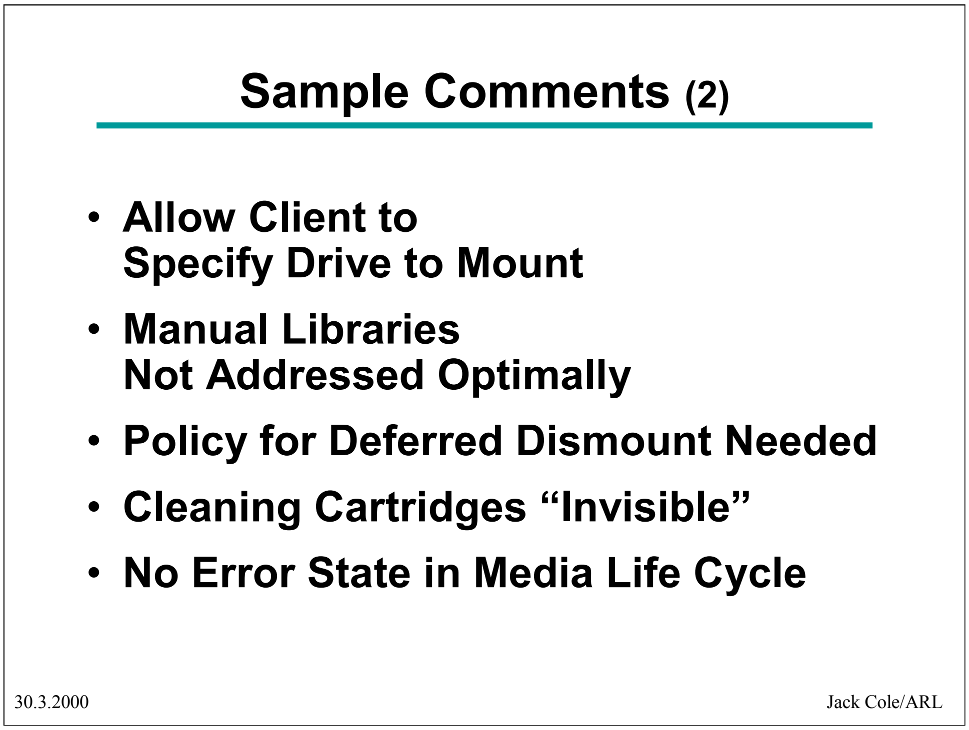
30.3.2000 Jack Cole/ARL

Counting Device Errors – Addition Made to Data Model

Non-removable Media – Yes, It Could. Non-Removable Media is just mounted less often ☺

Conformance Levels Are Needed – Deemed to be way to complex to accommodate. Besides, the IEEE wanted the ten MMS standards to be one, and SSSWG fought to keep ten for granularity of conformance and evolution.

Distributed Atomic Mounts – Can Be Effected With MMS As Is.



Client Specified Drive - The MATCH command may be used to select a drive.

Manual Libraries Insufficiently Addressed – Some clarification to the document made, but manual libraries are sufficiently addressed. Since the LM resolves to physical actions, these actions may be satisfied by a “protein robot” by changing the mode of specification of the action.

Deferred Dismount (and Lazy Un-mount) – This is not an external policy, but a function of the MM, and can be preset to be allowed/disallowed.

Cleaning Cartridges – These are identified in the DM, so they are not “invisible” when mounted or un-mounted.

No Error State in Media Life Cycle – An Error State has been added.

Sample Comments (3)

- **No Default Task Queue Priority**
- **Static Configuration File Should Not Be Required**
- **Need to Distinguish Disabled Online/Offline for Drives, Libraries**
- **Unload and Unmount Should Be Optional**
- **Asynchronous Notification Needed**

30.3.2000 Jack Cole/ARL

No Task Queue Priority – Added.

Static Configuration File – None required, document clarified.

Distinguish Disabled, Online/Offline WRT Drives, Libraries – Done.

Unload and Unmount should be optional – WG viewed this as a security hole if made optional.

Asynchronous Notification Needed (Events like "new media injected", "library went offline", etc) – Will be considered for next version.

Work Ahead for MMS

- Form Responses
  Make Modifications
- Re-Circulate to Balloters
  explain
- SGI OpenVault™ Patent Release
- Wrestle with Token Registry
- Submit May 5th, 2000 for Approval

30.3.2000 Jack Cole/ARL

Plans

- Finish MMS Standards
- Data Mover
- Guide for Storage System Design
- Tape Standards (P1563)
- SNIA Collaboration
- Collaboration with Metacomputing Community?

30.3.2000 Jack Cole/ARL

We are promiscuous in pursuit of standards, guides, recommended practices.

Our goal is to bring order to development of storage systems which now is a 'many splintered' thing ☺

Curtis Anderson/Turbolinux <canderson@turbolinux.com> proposed the Tape Standards.

Our collaboration with SNIA and DMTF has been to provide the MMS storage object definitions to CIM, on which WBEM is based. The effort is to align the CIM and MMS data models. In addition, SNIA will present to the IEEE CS SAB, probably in November 2000. And SNIA and SSSWG will hold a common meeting on occasion. Members of SSSWG have attended several SNIA/SML meetings, and held teleconferences. Bruce Haddon is to be credited with starting the collaboration, and Curtis Anderson with providing technical expertise on the IEEE data model.

The 'metacomputing community' includes the Remote Data Access WG of the GridForum (http://www.gridforum.org) and possibly the IEEE CS Task Force on Clustered Computing.

SNIA = Storage Networking Industry Association (http://www.snia.org)

DMTF = Distributed Management Task Force, Inc (http://www.dmtf.org)

CIM = Common Information Model (of DMTF)

WBEM = Web-based Enterprise Management

Tape Standards (P1563)

- **Portable Tape Driver Architecture**
- **Common Tape Drive Semantics**
- **Common Format for Data on Tape**

Contact Curtis Anderson if you want to get involved.
canderson@turbolinux.com

30.3.2000 Jack Cole/ARL

1563.1 - Recommended Practice for Portable Tape Driver Architecture
Provides a reference model for tape driver architectures that is portable across multiple operating system environments, fully featured, and high performance. A fully realized architecture that industry can base their implementations on that will reduce the effort required to support a new tape device on a given platform and thereby increase the available choice of drives on any given platform. This will benefit the application vendor and the end customer.

1563.2 - Common Tape Driver Semantics
Defines a common set of operations and semantics for access to tape drives across multiple operating systems platforms. Eases the task of porting and supporting applications that use tape storage across multiple operating system environments. This will enable application vendors to port to more platforms and thereby increase the end customer's available choices.

1563.3 - Common Format For Data On Tape
Defines a self-identifying format and record structure for the storage of data and meta-data on tapes, a structure that contains the key to understanding the format of the data stream as well the data itself. An analogue from the networking world would be the Document Type Definition (DTD) structure used to describe documents in XML (eXtended Markup Language). Enables data written by one application to be accessible by other applications without those applications having to know how each other encodes data written to tape.

Standards Not A Lock

- Living Documents
- Development Continual, Iterative
- Does Not Stop With Publication
- Non-Balloter Comments Important
- Goal → Widely Accepted Standards
- Balance Time Demands
  "Feature Creep"

30.3.2000 Jack Cole/ARL

General Observations, Motivations

- **“Software is a Mess”, Jim Gray**
  (hardware and media doing fine)
- **Real Focus**
  → Transparent Information Access
- **Transparency is hard to sell.**
- **Non-Traditional Approaches Needed**
  (capacity 100x/decade, throughput 10x/decade)

30.3.2000 Jack Cole/ARL

Jim Gray, ACM Turing Award winner this year for work in databases and the Microsoft Bay Area Research Center (BARC) Director has written two valuable articles. One is his acceptance speech for the Turing award, the other “Rules of Thumb in Data Engineering”. In these he recognizes what many do, that “software is a mess” (his quote). http://research.microsoft.com/~gray/

The traditional Server-Attached Disk (Device), or “SAD” (Garth Gibson/CMU) is transitioning toward Net-Attached Storage (NAS) by way of a pit stop in Storage Area Networks (SANs). The traditional SAD approach can’t serve the exploding needs of storage. While storage capacities are growing at 100X/decade, storage throughput is only growing at 10X/decade (Jim Gray, again).

Summary

- The First Storage System Standards in the World Are Nearing Publication.
- "Standards" Are Labor Intensive, But Bring Consistency Needed for Inter-operability, Distributability
- Collaborations Are Important: the Path to Technological Development Is a 'Many Splintered' Thing.

Thanks

- Balloters
- NASA
- WG Members

30.3.2000 Jack Cole/ARL

Contact

Jack Cole

jack.cole@ieee.org

(410) 278 - 9276

http://www.ssswg.org