University of Minnesota
Digital Technology Center
Intelligent Storage Consortium

David H.C. Du
du@cs.umn.edu

For more information on Intelligent Storage Consortium, please check
http://www.dtc.umn.edu/programs/DISC.htm
Overall Focus

- Emphasize the application of Advanced Storage Technologies
- A Balanced approach to research that includes:
  - Applications that need/use storage
  - Advanced and Emerging Storage Architectures
  - Advanced and Emerging Storage Technologies both software and hardware
  - Business Cases and aspects of the Storage industry
    - Market Trends
    - Product Directions
    - Effects of these disruptive technologies
    - Adoption rates
- Provide consortium members with not just technology research but a more complete and significant outcome
Initial Specific Focus

- Applied Object-based Storage Device (OSD) Active Storage (Datanomic Concept)
  - The Application of OSD and Active Storage Devices to different real-world problems (MPEG-21, data mirroring, iSCSI based data sharing and file systems)
  - Demonstrate the feasibility and advantages of OSD and Active Storage (modifying Lustre code)
  - Develop an understanding of the limitations of OSD and Active Storage from a theoretical and practical standpoint (many research issues)
- How OSD Active Storage addresses real-world problems facing the storage industry today and tomorrow
- Must make this a Win-Win value proposition for the University and Industry
University Participation

The Digital Technology Center
- Jim Licari and Tom Ruwart

Cross-disciplinary University participants
- Computer Science – Software Technology focus
  - David Du – Active disk, OSD, and Networking
  - Jon Weissman – Software systems, grid computing
  - Yongdae Kim – Computer and Network Security
  - Zhili Zhang – Networking and Internet Engineering
- Electrical Engineering – Hardware Technology focus
  - Ahmed Tewfik – Signal processing, wireless network
  - David Lilja – Computer Architecture, Distributed systems
- Carlson School of Business – Business focus
  - Bob Kauffman, Information and Decision Sciences
  - Alok Gupta
  - Gediminas Adomavicius
Potential Industrial Participation

- **EMC, StorageTek** – Storage Systems
- **Seagate, Qlogic** – Storage components
  - Development and demonstration of OSD
  - Development and demonstration of OSD Active Storage
- **Veritas** – Storage Software
- **Microsoft, Intel, IBM, Sun Micro** – I/O Subsystem, Network Storage System and Applications
- **Cisco** – Network Storage System
- **Storage Networking Industry Association (SNIA)**
  - Involvement with Technical Working Groups
  - Contact with leading-edge storage and networking companies
  - Help and guidance for useful projects
Changes - Technology

- Intelligent devices
- Storage capacity
- Ubiquitous connection
  - high speed IP network
  - wireless network
- Reducing cost
- Amount of data
Changes – User & Application

**User Demand**
- anytime, anywhere access
- heterogeneous/diversified user requirement
- growing needs for storage space

<table>
<thead>
<tr>
<th>Year</th>
<th>Storage Needs (MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1997</td>
<td>1,370</td>
</tr>
<tr>
<td>1999</td>
<td>3,500</td>
</tr>
<tr>
<td>2001</td>
<td>7,600</td>
</tr>
<tr>
<td>2003</td>
<td>14,350</td>
</tr>
</tbody>
</table>

storage needs from end users

Emerging data-intensive applications

---

Digital Technology Center
What is Datanomic?

- In an autonomic system, processes are performed automatically in response to internal causes and influences.
- Our proposed approach is to take a data-centric view of the system in which actions performed by, or on behalf of, a data object are taken at the direction of the data object itself. Therefore, such a system is called datanomic system.
Datianomic Computing
System behavior driven by characteristics of the data

- Automatic optimization to ever changing data requirements
  - Allocate resources according to increase in demand of the data
  - Transform data formats to support different applications

- Seamless data access from anywhere at anytime
  - Location and context aware access to data
  - Adaptive performance
  - Consistent view of each user’s data
  - Independent of platforms, operating systems, and data formats

- Exploit active object and intelligent disk
Basic Technologies and Assumptions of Datanomic

- Intelligent Storage Devices
- Object Storage Devices
  - Unique Object Id
  - Flat Object Directory + Strong Search Engine
  - Data + Meta-data + Attributes
- Network Attached Storage Devices
- Separation of Control and Data Paths
- Mobility of Data, Clients and Storage Devices
- Wireless and Intermittent Connectivity
Regional Organization

- Partition of regions: based on physical or logical affinity
- Single regional manager
- Clients
- Intelligent object-based storage devices
Regional component (1)

- **Regional Manager**
  - Object metadata management
  - Security related issues within/outside region
  - Naming service
  - Object replication, migration and consistency
  - Clients and OSD devices management (including mobile clients and devices)
Regional component (2)

- **Client**
  - End users or applications that access objects within a region
  - Client has a home region that stores important client information. The home region is allowed to move
  - Client can move freely among region
Regional component (3)

- **Intelligent Object-based Storage Devices**
  - OSD decides if a specific client is allowed to perform some operations
  - Perform data-directed operations specified by the object itself
Coordination within a region

- Regional Manager
- Desktop
- Laptop
- App Server
- IP Network
- Intelligent OSD
Coordination between regions
Research Issues

- Dynamic replica management
- Caching and consistency management
- Security
- Efficient and robust data paths
- Representative applications
Replication Overview

**Challenges**
- Large scale of RMs, devices, objects and clients
- Wide area environment
- User, data, and device mobility

**Solution**
- Dynamic replica management
- Cooperative replication
  - Utilize the intelligence of object and device

**Goal**
- Reduce the burden of regional manager
- Avoid hot spot
- Reduce the access delay and network traffic
- Provide continue access when RM is unavailable
How can active object help?

- user/application customization
- flexible functions
- Three ways
  - Object Metadata
  - Associated Method
  - Data Dependent Attributes

<table>
<thead>
<tr>
<th>Object GUID</th>
</tr>
</thead>
<tbody>
<tr>
<td>- system transparent</td>
</tr>
<tr>
<td>- user-directed itinerary</td>
</tr>
<tr>
<td>- pre-captured access pattern</td>
</tr>
<tr>
<td>- lifetime</td>
</tr>
<tr>
<td>- component objects</td>
</tr>
<tr>
<td>- copy(source, dest, format, bytes)</td>
</tr>
</tbody>
</table>
How can intelligent device help?

**What does intelligent device know?**
- Object read/write traffic
- Real-time device load

**What can intelligent device do?**
- inside region object replication
  - when and where(device) to replicate
  - discover neighbor devices
  - load balance traffic to neighbors
- guide cross region object replication
  - when and where(region) to replicate
- provide continue access when RM is down
- alleviate RM burden (more scalable inside one region)
Role of RM

- Speculative replication based on “open” request
- Cross region replication
- Backup or mirroring
- Distributed location service
- Quota Control
Consistency Management Overview

- Recognize replicas as the same object
- Allow automatic duplication, deletion, or migration of objects
- Support concurrent access of the same logical object from nearby replicas
- Allow caching of objects at multiple levels
- Overcome intermittent connections
Levels of Consistency

- Objects indicate desired consistency level
- **Strong consistency**
  - global consistency at any moment
  - using *server locking*
- **Weak consistency**
  - allow temporary inconsistency
  - using *optimistic concurrency control*
- **Balancing in-between**
  - Could allow one object to split into two versions
Partition of Consistency Management Functions

Regional Managers + Intelligent Storage Devices

Goals:
- reducing overheads
- enhancing concurrency

Approaches:
- fine granularity smaller than data objects
- hierarchical management over RM/ISD hierarchy
Differential Updates of Objects

- Optimistic concurrency control generates co-existing multiple versions of objects
- Update history has to be maintained until final reconciliation
- Full copy of individual versions consume too much storage
- Exploiting incremental/differential updates
  - Existing scheme for text file and database app.
  - Seeking more general support for various objects
Distributed Committing

- Updates of compound object may involve several referenced objects.
- Such updates have to be performed in a transaction way.
- Such updates also have to support versions such that rollback is possible.
- Determine global serializability of distributed updates.
Consistency Management under Mobility

Mobile Entities:
- Users w/ high mobility
- Active objects w/ less mobility
- Intelligent storage devices w/ less mobility

Different requirements
- Users require quick hand-off when moving across adjacent regions
- Active objects & Intelligent storage devices requires least interruption of other services during the moving
Security Issues

- Typically file managers are involved in all security-related functions
  - Issuing capability, access control, revocation
  - Excessive load on the file managers
  - Central point of failure
  - Attractive target of DoS attacks

- Devices are directly attached to the network
  - Exposed to various potential attacks
  - DoS, masquerading, replaying, other active/passive attacks
Security Issues

- Direct interaction between client and devices
  - Client needs a capability or credential from the RM
  - Object Mobility
  - Compound objects
  - Devices need to perform additional functions
    - Authentication, access control

- Traffic analysis
  - Traffic trends or identity of the communicating party can reveal some sensitive information
    - Merger of two companies, business trends of ISP
  - Providing anonymity of traffic is important
Authentication

General authentication
- Client – RM, client- device, device – RM, RM – RM

Message authentication
- Messages sent from and between RMs
- Responses from devices

Non-repudiation
- Preventing RM, device, client from denying their operations or commands
Access Control

- Credential based
  - ACL on the RM
  - RM issues a credential which contains the access rights for the client
  - This credential by the client to probe to the device his access rights
  - Credential usually short lived, frequent requests to RM
  - RM has to be online for the entire system to work

- Identity based
  - ACL associated with the object
  - Client receives an identity certificate from the RM
  - Preferred if the object is mobile
  - Revocation difficult, changes in client’s rights results in large number of updates
Access Control

- Compound objects
  - The component objects could be independent objects with different access rights than the main object
  - Could be located on different device or region
Confidentiality and Integrity

- Insecure Internet communication
  - Encryption on link necessary
- Person getting physical access to the device can read and modify the objects
  - System administrator, device thefts
  - End to End object level encryption may be required (optional)
- Ensuring integrity of the traffic on the link, cache etc. is important
- Compound objects
  - How to encrypt?
  - Accessing one object might require more than one keys
Key Management and Keyword Search

- **Key management**
  - Merging different groups
  - Revocation of the user
  - Encrypted objects can reside on the device for a long time

- **Some parts of meta-data will be encrypted**
  - How to support search operations using this meta-data
  - Decrypting every time will be expensive

- **If client wants to retrieve objects that contain certain keywords**
  - RM will have to decrypt all of the relevant objects
  - However, RM cannot always be trusted (E2E)
  - Decryption of the objects will be expensive and unnecessary
Data Path Requirements

- High performance
  - Reduce the protocol and system software overhead
- Quality of service
  - Object data delivery adapts to network condition
- Mobile client
  - Gracefully handle the intermittent connection
- Wireless connection
  - Optimize the frequency, power level allocation
- Robustness
  - Handle failure in the data path (e.g. network failure, end system crash, etc.)
Software Overhead in Data Path

- Software overhead of TCP/IP in end system is two high:
  - Multiple memory copies
  - Protocol processing overhead
  - Context switching
  - Interrupt servicing

- The object device has limited processing power
  - CPU bound
  - Memory bound
Reducing Software Overhead

- **OSD/Secure RDMA**
  - Use RDMA to achieve zero memory copy
  - Leverage RDMA over TCP/IP
  - Incorporate security with RDMA mechanism
  - Integrate RDMA with TOE
  - Exploit the caching for different clients
  - Develop a RDMA access interface
OSD/Secure RDMA Architecture

OSD Client

Application

Buffers

OSD VIPL

VI NIC driver

NIC

IP network

OSD Device

OSD controller

Buffers

OSD VIPL

Object Manager

Disk Driver
Quality of Service

- Active object is a special application
  - may have its own QoS requirement
    - Latency variance
    - bandwidth

- Underlying network condition unpredictable
  - TCP is a best-effort protocol, no guarantee of BW
  - Traffic can be bursty and varied
  - Congestion may occur

- How to provide the QoS for object in datanomic?
- How to adapt the QoS smoothly when the network condition degrades?
- How to allot and schedule the BW in a OSD device?
Adaptive QoS control

- Collect status info of substrate network
- Exchange QoS info between application and substrate network
- Exploit application “hint”
- Tailor the application level data transmission based on the underlying network behavior
- Network level transmission also adjusts to the above application requirements
Mobility

- Mobile clients (PDA, wireless and mobile users)
- Mobile objects
- Mobile storage devices

Challenges

- Connection can be intermittent
- An entity can dynamically join and leave a region
- A transition may happen in the process of data transmission or critical operation
Wireless Connection

- Limited resource in wireless connection
  - Bandwidth, error rate, power consumption
- Effective resource management
  - Optimize frequency allocation, power levels, and beamforming of access points
  - Adapt to changing traffic patterns and QoS requirements
  - Exploit multiple channels
Robustness

- Possible component failures in datanomic system
  - Network failure
  - Storage device failure
  - Client failure
  - Regional manager failure

- Maintain robustness in the face of failures
  - Data consistency
  - Graceful degradation
  - Roll-back mechanisms
Representative Applications

- MPEG-21-compliant applications
  - Meeting users’ heterogeneous capabilities, dynamic bandwidth changing, & display resolution by:
    - Replication of objects and compound objects in different formats and resolution.
    - Layered Multimedia representations.
    - On-the-fly transcoding.
  - Media personalization and Content-based search techniques.
Representative Applications (continue)

- Backup strategies (special case of replication) that operate at multiple time scales.

- Potential application in the field of the medical research and data mining:
  - Using the “methods” stored together with active data objects to generate requested statistics or discovering undefined correlations based on a large number of patient records without effecting the confidentiality of patient records.