Introspective Replica Management in OceanStore

Dennis Geels
Problem Description
Replica Data Structure

- Heartbeats
  - Signed (name → version) mapping
- Version Metadata
  - Chain of Metadata blocks for recent versions
- Resource Accounting
  - % of each version stored locally
  - Access frequency
  - Payment/contract info (eventually)
Read Algorithm and Latencies

1) Open Replica
   – One Tapestry lookup round trip
2) Determine version
   – One round trip to parent replica
3) Read BTree
   – Round trip to parent replica per level of BTree, or
   – Forward request to parent (one large transfer)
Why Bother?

• Problems caused by poor Replica Management:
  1) High access latency
  2) Sensitivity to Network Failure
  3) Poor load sharing
Solution

1) Keep Replica near clients
   • Makes OceanStore behave like a local area file system
   • Improves latency, robustness, load sharing

2) Do work ahead of time
   • Avoids initial wide-area hit
How to Prefetch?

• **Algorithm:**
  1) **Cluster files**
     - Manually / automatically
  2) **Detect access to Cluster**
     - Possibly predict use via temporal patterns?
  3) **Prefetch Cluster in stages**
     - Initiate read process
     - More effort expended on most likely files
Who Prefetches?

- Client machines
  - Access to user activity stream
  - Limited resources
- Proxy machine
  - Off-node helper with extra storage, better network connectivity
Where do we find proxies?

• Choose machine along search path
  – Saves work in following Tapestry lookup
• Keep proxy on retainer
  – Quick lookup, simple contracts
• Bulletin board
  – Post requests to third-party server
  – Any available proxy responds
Example: Using Search Path

- Clients dispatch prefetch requests along normal read path
- Can use TTL to place more important replicas nearer
Example: Proxies on Retainer

- Client machines maintain contact with small number of local proxies
- Can find resources off lookup path
- Direct communication, simple reputation management
Example: Using Bulletin Board

- Third party accepts requests for various services
- Local proxies with available resources receive updates
- Very flexible, general infrastructure support for proxies
Future Work

- Smart decisions when traffic dominated by updates
  - Fewer replicas often better
- Large scale testing, simulation
Conclusion

Introspective Replica Management may negate the problems of wide-area storage. OceanStore will (hopefully) feel like a local network file system.
Acknowledgements

• My Wife, Cheryl