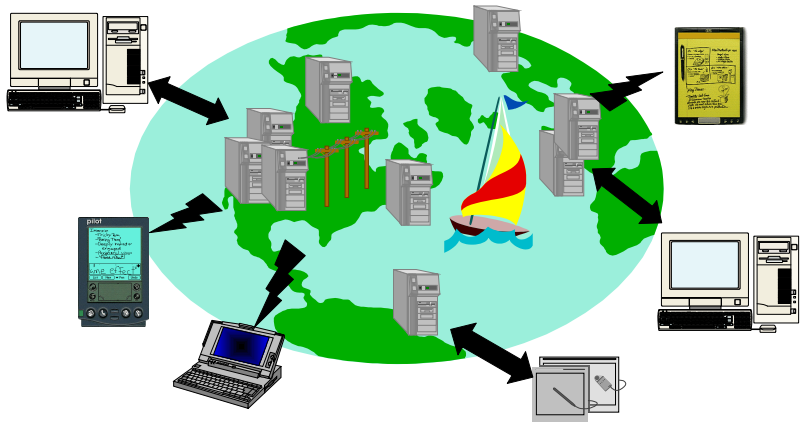
Option 2:
The Oceanic Data Utility:
Global-Scale Persistent Storage

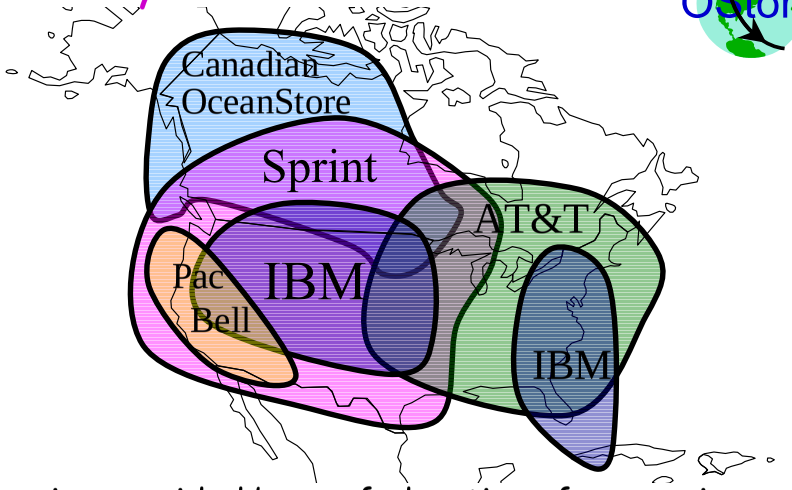


John Kubiatowicz

Ubiquitous Devices ⇒ Ubiquitous Storage

- Consumers of data move, change from one device to another, work in cafes, cars, airplanes, the office, etc.
- Properties *REQUIRED* for Endeavour storage substrate:
  - Strong Security: data must be encrypted whenever in the infrastructure; resistance to monitoring
  - Coherence: too much data for naïve users to keep coherent "by hand"
  - Automatic replica management and optimization: huge quantities of data cannot be managed manually
  - Simple and automatic recovery from disasters: probability of failure increases with size of system
  - Utility model: world-scale system requires cooperation across administrative boundaries

Utility-based Infrastructure



- Service provided by confederation of companies
  - Monthly fee paid to one service provider
  - Companies buy and sell capacity from each other

State of the Art?

- Widely deployed systems: NFS, AFS (/DFS)
  - Single "regions" of failure, *caching only at endpoints*
  - ClearText exposed at various levels of system
  - Compromised server ⇒ all data on server compromised
- Mobile computing community: Coda, Ficus, Bayou
  - Small scale, fixed coherence mechanism
  - Not optimized to take advantage of high-bandwidth connections between server components
  - ClearText also exposed at various levels of system
- Web caching community: Inktomi, Akamai
  - Specialized, incremental solutions
  - Caching along client/server path, various bottlenecks
- Database Community:
  - Interfaces not usable by legacy applications
  - ACID update semantics not always appropriate

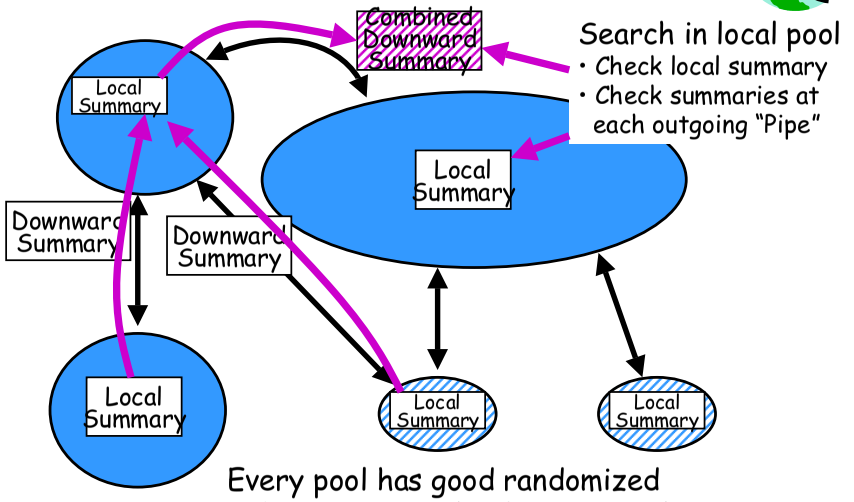
OceanStore Assumptions

- Untrusted Infrastructure:
  - Infrastructure is comprised of untrusted components
  - Only cyphertext within the infrastructure
  - Must be careful to avoid leaking information
- Mostly Well-Connected:
  - Data producers and consumers are connected to a high-bandwidth network most of the time
  - Exploit mechanism such as multicast for quicker consistency between replicas
- Promiscuous Caching:
  - Data may be cached anywhere, anytime
  - Global optimization through tacit information collection
- Operations Interface with Conflict Resolution:
  - Applications employ an operations-oriented interface, rather than a file-systems interface
  - Coherence is centered around conflict resolution

OceanStore Technologies I: Naming and Data Location

- Requirements:
  - Find nearby data without global communication
  - Don't get in way of rapid relocation of data
  - Search should reflect locality and network efficiency
  - System-level names should help to authenticate data
- OceanStore Technology:
  - Underlying namespace is flat and built from cryptographic signatures (160-bit SHA-1)
  - Data location is a form of gradient-search of local pools of data (use of attenuated Bloom-filters)
  - Fallback to global, "exact" indexing structure in case data not found with local search

Cascaded-Pools Hierarchy



Every pool has good randomized index structure (such as Treaps)

OceanStore Technologies II: High-Availability and Disaster Recovery

- Requirements:
  - Handle diverse, unstable participants in OceanStore
  - Eliminate backup as independent (and fallible) technology
  - Flexible "disaster recovery" for everyone
- OceanStore Technologies:
  - Use of erasure-codes (Tornado codes) to provide stable storage for archival copies and snapshots of live data
  - Mobile replicas are self-contained centers for logging and conflict resolution
  - Version-based update for painless recovery
  - Redundancy exploited to tolerate variation of performance from network servers (RIVERS)

OceanStore Technologies III: Introspective Monitoring and Optimization

- Requirements:
  - *Reasonable* job on a global-scale optimization problem
  - Take advantage of locality whenever possible
  - Sensitivity to limited storage and bandwidth at endpoints
  - Stability in chaotic environment
- OceanStore Technologies:
  - Introspective Monitoring and analysis of relationships:
    - between different pieces of data
    - between users of a given piece of data
  - Rearrangement of data in response to monitoring:
    - Economic models with analogies to simulated annealing
  - Sub problem of Tacit Information Analysis (option 5)

OceanStore Technologies IV: Rapid Update in an Untrusted Infrastructure

- Requirements:
  - Scalable coherence mechanism which provides performance even though replicas widely separated
  - Operate directly on encrypted data
  - Updates should not reveal info to untrusted servers
- OceanStore Technologies:
  - Operations-based interface using conflict resolution
  - Use of incremental cryptographic techniques: No time to decrypt/update/re-encrypt
  - Use of oblivious function techniques to perform this update (fallback to secure hardware in general case)
  - Use of automatic techniques to verify security protocols

Two-Phase Implementation:

- Year I: Read-Mostly Prototype
  - Construction of data location facility
  - Initial introspective gathering of tacit info and adaptation
  - Initial archival techniques (use of erasure codes)
  - Unix file-system interface under Linux ("legacy apps")
- Year III: Full Prototype
  - Final conflict resolution and encryption techniques
  - More sophisticated tacit info gathering and rearrangement
  - Final object interface and integration with Endeavour applications
  - Wide-scale deployment via NTON and Internet-2