AVAMAR

Managing Dynamic Archives

Dr. Kevin C. Daly, CEO Avamar Technologies

THEN NOW

… there’s plenty to do

- According to UCB’s School of Information Management & Systems
  - We create 5,500PB of new information annually
    - 3,500PB (64%) is digital
      - 2,000PB (37%) on hard disk
  - The Internet represents an information flow of 500PB annually – all of it digital
  - The telephone represents an information flow of 17,300PB annually – mostly analog, but rapidly becoming digital
- I would add
  - Internal data networks (LANs & SANs) represent an additional information flow of 400,000PB annually

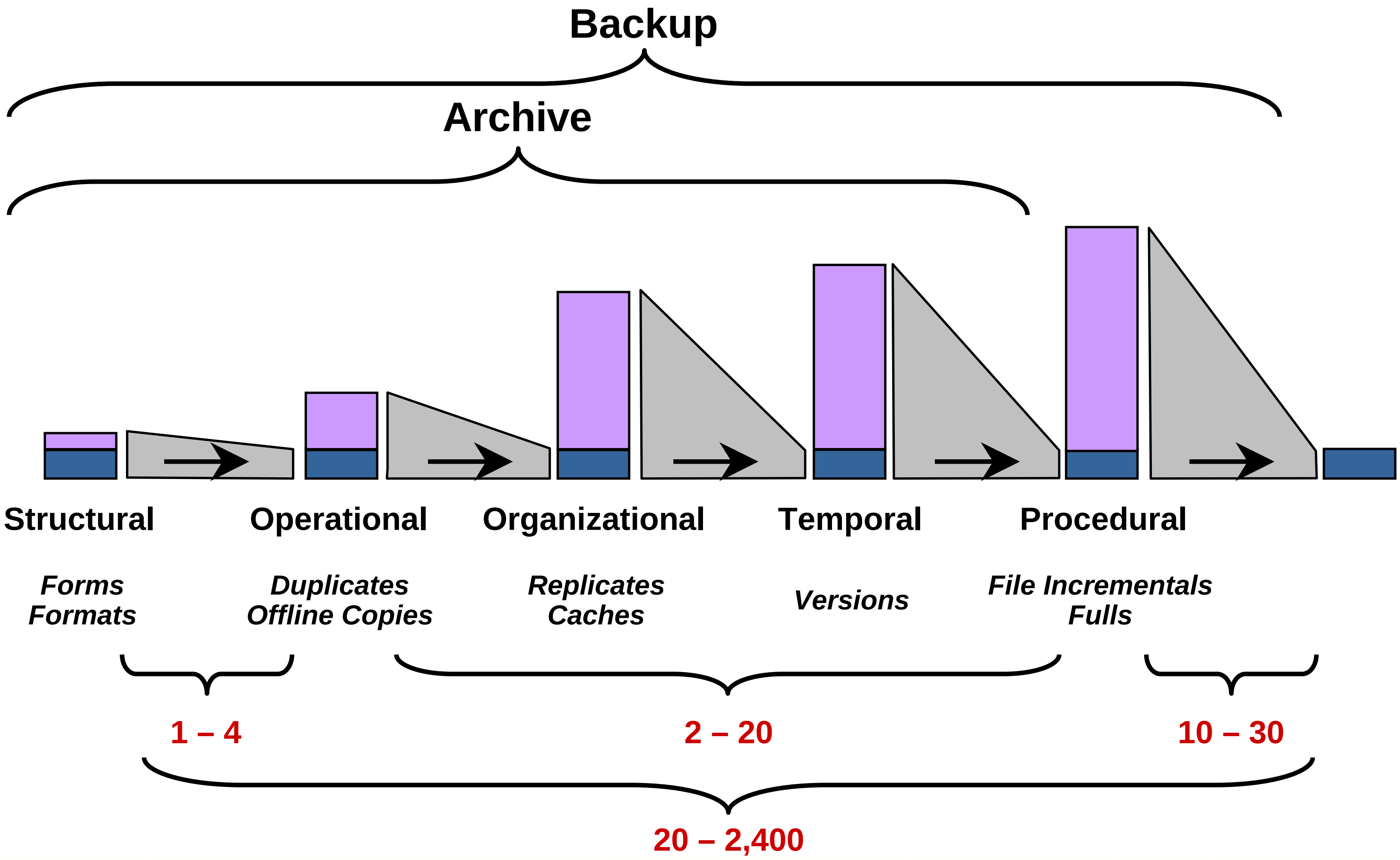
Points to Ponder

- The Information Density of data at rest is low: <<10% (i.e. Data Redundancy is high)
- The Information Density of information flows is even lower: <<1%
- Any move toward archiving information flows will increase the opportunity space for archiving by more than an order of magnitude

---

- Factoid: current disk production is 20,000PB annually on a base of 40,000PB with 25% going into shared environments

Sources of Data Redundancy



Thesis

The technical and economic state of magnetic disk technology, when combined with several key data management and file system technologies permit the practical use of disks for a number of applications that had been considered traditional tape applications

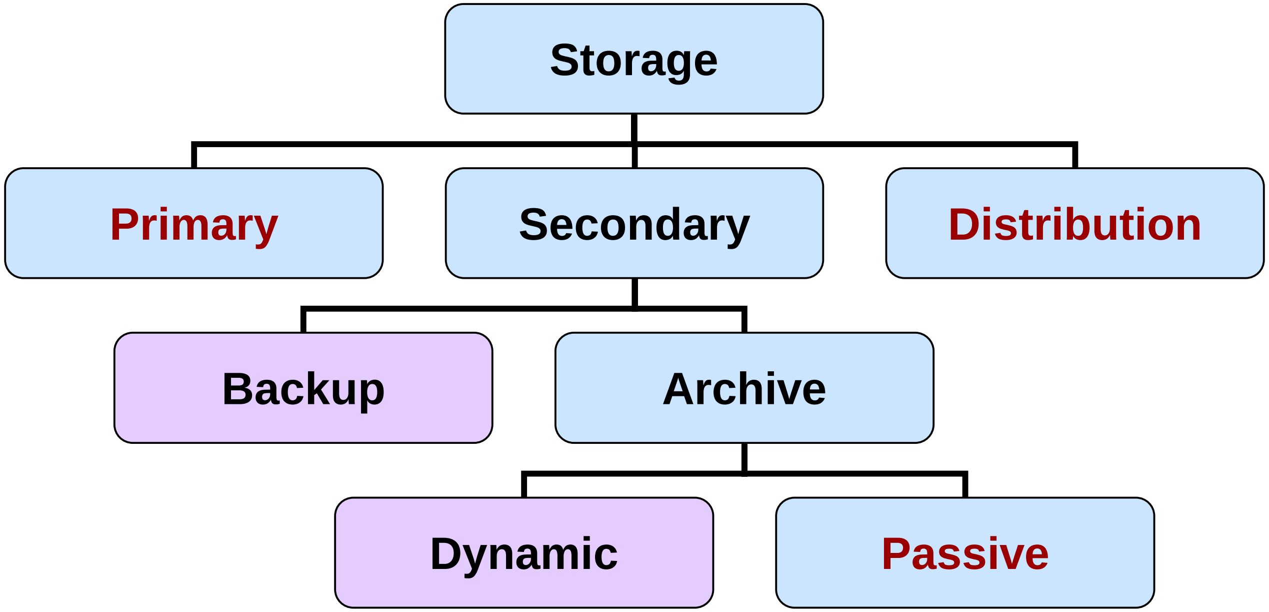
Backup

Dynamic Archive

The continuing cost trajectories of disks will accelerate the adoption of disk-based systems over the next five years, and data efficiency will be the key differentiator among the approaches

AVAMAR

Storage Taxonomy

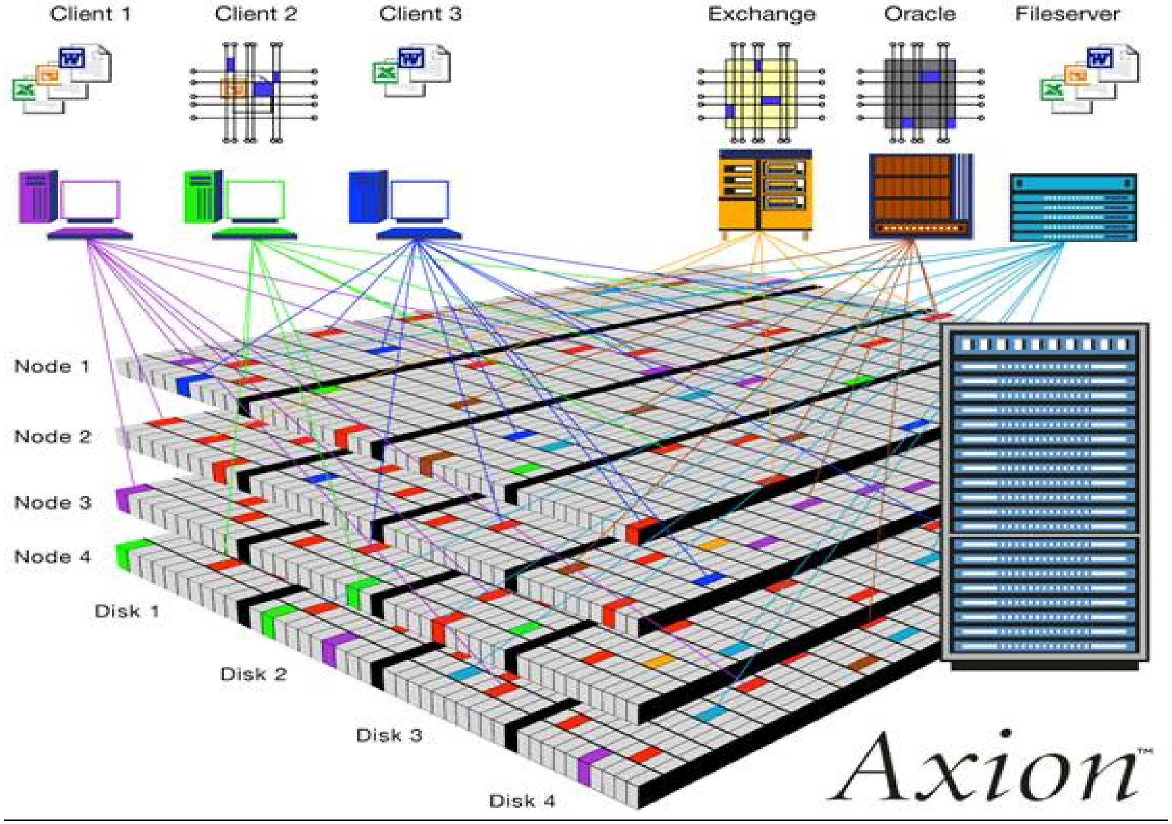


Secondary Storage

- Home for disruptive technologies:
  - Storage networking
  - Content Addressed Storage (CAS) File Systems
  - sATA Drives & Silicon sATA Controllers
  - Redundant Array of Independent Nodes (RAIN)
- Battleground of the *media wars*
- Includes significant software value
  - Typically 50% – 100 % of hardware value
  - Often from 3rd parties – e.g. backup,
    e-mail archive

Content Addressed Storage (CAS)

- All data objects given a *Content Address*
- Content Address specifies exact storage location for data object
- Content address can be used to locate and restore data objects
- Content addresses assure load-balanced reads and writes
- Unique objects are stored only once and are shared among all clients



Media Wars: Why Tape?

- Capacity (~2:1 advantage)
  - Tape: 150 – 500 GB
  - Disk: 80 – 250 GB
- Density (~3:1 advantage)
  - Tape: 10 – 30 GB/in^3
  - Disk: 3 – 10 GB/in^3
- Cost (~2:1 advantage)
  - Tape: $1/GB
  - Disk: $2/GB
- Export (~10:1 advantage)
  - Tape: >5 TB in 24 hrs
  - Disk: ~0.5 TB in 24 hrs
- Passive Archive
  - Tape: Yes (controlled environment)
  - Disk: No

Media Wars: Why Disk?

- Access Latency (~10,000:1 advantage)
  - Tape: ~100 sec
  - Disk: ~10 msec
- Rate Range (>250:1 advantage)
  - Tape: 4:1
  - Disk: >1,000:1
- Redundancy
  - Tape: Replication only
  - Disk: RAID
- Active Media Life (10:1 advantage)
  - Tape: Months
  - Disk: Years
- Integrity Validation (>300:1 advantage)
  - Tape: Yearly (at most)
  - Disk: Daily (typically)

System Software Value

- Effective use of ATA drives in enterprise environments requires unique support at a system level
  - Integrity: tolerance of *seek errors*
  - Reliability: tolerance of AFR 2x to 3x that of SCSI/FC drives
  - Efficiency: accommodation of high *data-under-a-head* ratios
- Properly implemented, CAS supports:
  - Integrity: device-independent read error detection
  - Reliability: RAID-class parity protection
  - Efficiency: distribution of read/write activity among available drives

---

- *Prediction*: Outside of CAS, ATA drives will find limited applications in enterprise environments

Summary & Conclusions

- Archives will become more dynamic
  - Because they *can*: sATA drives & CAS File Systems
  - Because they *must*: integration into the storage hierarchy for reference data, compliance, etc.
- Information Density (removal of data redundancy) is critical for keeping up with demand growth
- Hardware developments are important, but software developments are even more important
  - IDC estimates that the rate of growth of storage software revenue over the next five years will be three times the rate of growth of hardware revenue