

Engineering Issues in Preserving Large Databases



David S. H. Rosenthal

LOCKSS Program
Stanford University Libraries

<http://www.lockss.org/>

© 2007 David S. H. Rosenthal

LOTS OF COPIES KEEP STUFF SAFE

Who Am I?



- 21 years in Silicon Valley – 3-for-3 record
 - Distinguished Engineer @ Sun
 - Employee #4 @ Nvidia
 - Helped start Vitria Technology
- 28 years Unix system administration
 - Since 6th Edition on PDP-11/45
 - Built & ran Nvidia's 1st chip simulation farm
 - Built & ran Vitria's 1st enterprise middleware test lab
- Chief Scientist, LOCKSS @ Stanford Library

Science & Engineering



- I'm "Chief Scientist" but really an engineer
 - Science is about facts & ideal systems
 - Engineering is about money & real systems
 - *Real systems are inevitably imperfect*
- Need cost-performance tradeoffs:
 - Right now we only have rules of thumb
- This talk is advocating a *research agenda*
 - E-science databases are big and therefore expensive
 - Good economic decisions will matter a lot
 - They need to be based on good science and good data

Preservation *is* Fault Tolerance



- Bits *can* be copied perfectly
 - This doesn't mean they always *will* be copied perfectly
 - Perfect preservation is neither *guaranteed* nor *free*
 - In fact, at a large enough scale, it is *impossible*
 - How much loss can we tolerate?
- Everything that can possibly go wrong, will
 - How often will things go wrong?
 - How well will we tolerate things going wrong?
- We want better, more affordable preservation
 - Must *predict, measure & trade-off* cost and performance

Threat Model



- Media failure
- Hardware failure
- Software failure
- Network failure
- Obsolescence
- Natural Disaster
- Operator error
- External Attack
- Insider Attack
- Economic Failure
- Organization Failure

Rules of Thumb



- Safer data but higher cost from:
 - More replicas
 - BFT: $3f+1$ replicas survive f simultaneous faults
 - More independent replicas
 - Less correlation between faults, therefore
 - Fewer simultaneous faults
 - More frequent audits of replicas
 - Shorter lifetime of latent faults, therefore
 - Lower probability of coinciding faults

LOTS OF COPIES KEEP STUFF SAFE

How Well Must We Preserve?



- Keep a petabyte for a century
 - With 50% chance of remaining completely undamaged
- Consider each bit decaying independently
 - Analogy with radioactive decay
- That's a bit half-life of 10^{18} years
 - One hundred million times the age of the universe
- That's a very demanding requirement
 - Hard to measure
 - Even *very* unlikely faults will matter a lot

How Likely Are The Threats?



Examples:

- **Hardware**
 - Schroeder 2007
 - Pinheiro 2007
- **Software**
 - Prabhakaran 2005
 - Yang 2006
- **Operator Error**
 - "Most important cause of data loss"
- **Internal Attack**
 - Secret Service report
 - Under-reported
- **External Attack**
 - Software mono-culture
 - Flash worm

Example: Disks



- Manufacturers specifications:
 - 10^6 hours MTTF
 - 10^{-14} unrecoverable bit error rate
- Schroeder & Pinheiro FAST '07 papers:
 - Field replacement rate 2-20 time MTTF
 - No "bathtub curve" of early failures
 - Enterprise disks 10x expensive, no more reliable
 - No correlation between temperature & failure
 - Significant autocorrelation – very bad for RAID
 - Significant long-range correlation
 - SMART data logging not useful for failure prediction

Example: Software



File system code is carefully written & tested:

- Iron File System (Prabhakaran 2005):
 - Fault injection using pseudo-driver below file system
 - Bugs and inconsistencies in ext3, JFS, ReiserFS, NTFS
- FiSC (Yang 2006):
 - Model checking of file system code
 - 33 severe bugs in ext3, JFS, ReiserFS, XFS
 - Could destroy / in each file system
- Take away message:
 - The more you look, the more you find

Example: Insider Attack



- Political interference (Hansen 2007):
 - 2006 Earth Science budget *retroactively* reduced 20%
 - "One way to avoid bad news: stop the measurements!"
 - Suppose the data itself turned out to be "inconvenient" ...
- Independent replicas essential
 - In different jurisdictions

Fundamental Problem



- Perfect preservation - not at any price
 - Threats too prevalent, diverse, poorly understood,
 - Real systems are inevitably imperfect
- How imperfect is adequate?
 - How much will it cost?
- How adequate is what we can afford now?
 - Won't know unless we can measure performance
- Kaizen: improve cost-performance thru time
 - Need preservation benchmarks to drive market
 - Learn from incidents c.f. NASA's ASRS

Measuring Preservation



- Look at an exabyte of data for a year?
 - See ~5 bit flips? How sure are you?
 - Not affordable!
 - Doesn't match threat model!
- Fault injection?
 - How to inject realistic threats?
 - Insider attacks?
 - Natural disasters?
 - Have to inject huge numbers of faults!
- Other ideas?

Dynamic Economics



- Varying, uncertain *time value of money*
 - Postpone replication, but adds to risk
- Rapid, predictable decrease in *cost-per-byte*
 - Postpone replication, but adds to risk
- Rapid increase in *total demand* for storage
 - Replicate now, before competitors grab funding
- Varying, uncertain *future funding probability*
 - Repeated economic triage inevitable
- Endowment is the only safe mechanism

Service Level Agreements



- Create dataset, endow it, hand off to service:
 - Service level agreement to specify quality of preservation
 - Otherwise market captured by Potemkin services
- How to write the agreement?
 - How can we specify performance we can't measure?
- How to audit compliance with agreement?
 - LOCKSS: mutual audit protocols *between replicas*
 - Other ideas?

Transfer Of Custody



- Liability Disclaimers are endemic:
 - AMAZON DOES NOT WARRANT THAT AMAZON WEB SERVICES ... WILL BE ACCESSIBLE ON A PERMANENT BASIS OR WITHOUT INTERRUPTION OR THAT THE DATA YOU STORE IN ANY SERVICE ACCOUNT WILL NOT BE LOST OR DAMAGED
- Liability Disclaimers are viral:
 - You can't accept liability for your suppliers' products
 - Disclaiming lowers your competitors' costs
- Transfer of custody without liability:
 - Can it be meaningful?

Research Agenda



- Better data on incidence of threats
 - disk behavior, bugs, operator errors, attacks, ...
- Better algorithms & architectures
 - a "better than BFT" model?
 - "better than TPM" hardware support for preservation?
 - highly independent replica architectures?
- Better cost-performance models
 - Define, measure "performance" of preservation systems?
 - Taking decisions with dynamic costs & performances?
 - Transferring custody of data vs. liability disclaimers?

Credits



- LOCKSS Engineering Team (since 1998)
 - Tom Lipkis, Tom Robertson, Seth Morabito, Thib G-C.
- LOCKSS Research Team (since 2001)
 - Best Paper @ SOSP2003
 - Mary Baker, Mehul Shah & colleagues @ HP Labs
 - Mema Roussopoulos & students @ Harvard CS
 - Petros Maniatis & interns @ Intel Research Berkeley
- Vicky Reich, and funding from
 - NSF, Mellon, libraries, LoC, publishers, Sun, ...