How Green Is Digital Preservation?

David S. H. Rosenthal

LOCKSS Program
Stanford University Libraries
http://www.lockss.org/

LOTS OF COPIES KEEP STUFF SAFE

Not Green Right Now

Monthly Data Center Cost	
Hosts (divided over 3 year life)	$18,416.67
Bandwidth	$8,395.06
Space/Power/Cooling	$41,045.93
Network (divided over 3 year life)	$2,222.22
Datacenter	**$70,079.88**

Vijay Gill, Google, via The Register

Change Is On The Way

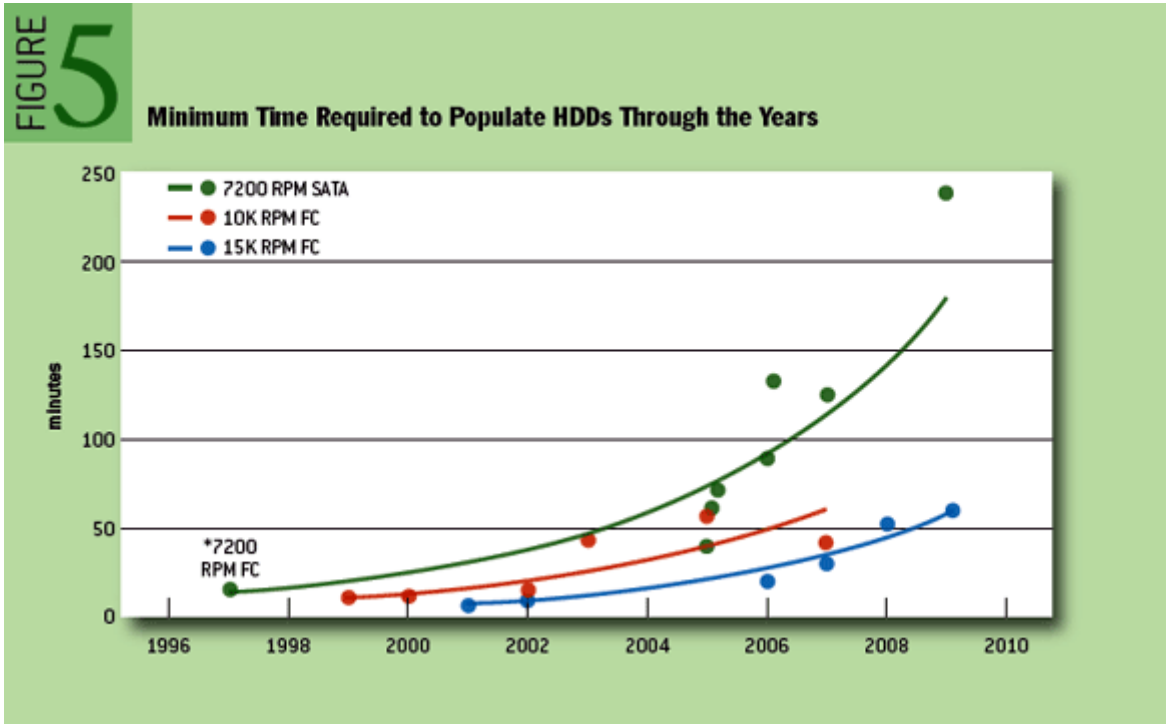
"FAWN couples low-power embedded CPUs to small amounts of local flash storage, and balances computation and I/O capabilities to enable efficient, massively parallel access to data. ... FAWN clusters can handle roughly 350 key-value *queries per Joule* of energy – two orders of magnitude more than a disk-based system"

“... small-object random-access workloads are ... ill-served by conventional clusters ... two 2GB DIMMs consume as much energy as a 1TB disk. The power draw of these clusters is ... up to 50% of the three-year total cost of owning a computer.”

David Andersen et al. "FAWN A Fast Array of Wimpy Nodes", SOSP, Oct. 2009

LOTS OF COPIES KEEP STUFF SAFE

Performance *And* Economy?



- Data on disk moving away from the CPU
- Why does this matter for preservation?
  - Fixity checks need to read the whole drive repeatedly

Archival Media

- Paper as archival medium
  - Survives benign neglect for a long time
- Magnetic archival media
  - Survives benign neglect but only for a few years
- Solid state memory as archival medium
  - Robust packaging, zero power for retention
  - Interfaces – Ethernet, TCP/IP – highly stable
  - Will survive benign neglect just fine

Architectural Implication

- Current systems create & store metadata
  - Indexes, format databases, ...
  - Run query once & record result
  - Because it's too expensive to look at the data each time
- Risk that stored metadata is inaccurate
  - As system gets bigger, risk increases
- FAWN architecture = cheap to look at data
  - Run the query each time, so result reflects current reality