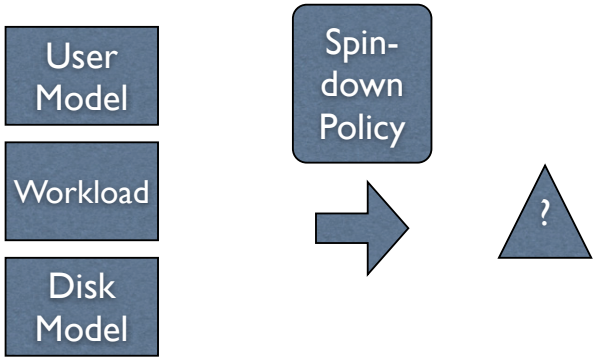
Storage 1: To Sleep or Slow, that is the ?

David Andersen
Low-Power Computing
Carnegie Mellon University



Disk Model

- Spin up time, spin down time, idle power, active power, spin up/down power
- One important derived metric: The break-even threshold.
  - Power(spin down, idle, spin up) < power(active) over a period of time
- (iff spin up power > active power)

User model key

- How much delay does a user mind?
  - Traditional metrics measured in ~10-300ms range...
  - Clearly does not apply to disk (1-10s)
- Managing expectation:
  - If away from computer for a while, delay OK;
  - If actively using? Delay annoying.
- Seems reasonable - no user studies that I know of.

Random HCI aside

- Stokes Ph.D. thesis 1991:
  - The right amount of delay (on the order of seconds, even) can *help* by giving the user some time to think.
    - "Delay must be appropriate for the cognitive task"
- But CS folks don't like delay, so we'll ignore this and assume it's all bad. :-) In fairness, you can always add in delay later -- it's hard to remove it.
- And opening mail reader after lunch - low cognitive load.

Policy

- AIAD - additive increase, additive decrease
  MIMD - multiplicative inc, mult. dec
  - Exhaustive-ish exploration of parameter space
- (One wonders why they didn't try the obvious AIMD, but that's my networking self speaking) - others have since tried it - it works a bit better, but ...

Bottom Line?

- On laptops, sleeping is very good, but how much?
- It's a tradeoff
- Using *an* adaptive policy modestly better than fixed - but *which* policy is best? No clear winner.
- Later studies: Some policies can modestly outperform fixed (and on their study, the adaptive algos from today didn't work well), but...

Fairly dismal results

	Power	Bad spindowns
Perfect	1.64	0 wrong
On	3.48	0 wrong
T = 30	2.05	18 wrong
Best-ish	1.94	15 Wrong

Lu et al, "Quantitative Comparison of Power Management Algorithms"

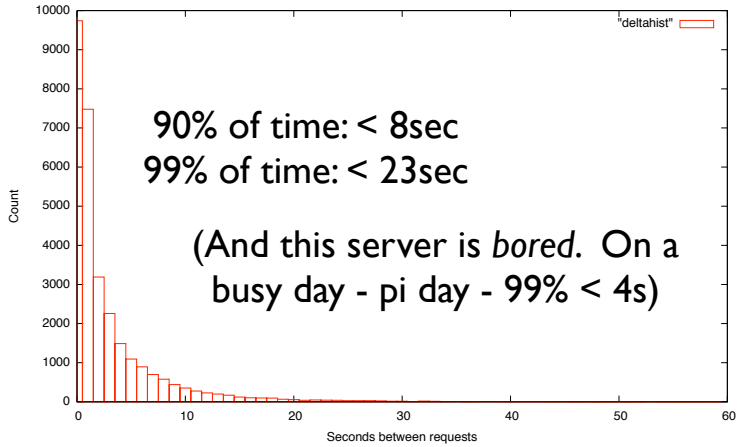
The good news

- Performance and power are *not* incompatible
- 32KB of NVRAM (paper calls it SRAM, but that's just a choice - you can do battery-backed SRAM or DRAM; SRAM doesn't need refresh, but $, as discussed) can avoid a lot of write-mandated wakeups
- A few MB of DRAM can help avoid some wakeups
- But this only works if small working set!
- ex: DEC SRC MP3 player - read-ahead, aggressive disk sleep to get power - but very predictable workload

Servers?

- Fairly constant activity; load varies over day
- eg from my own (tiny!) web server
  - served 30,054 requests yesterday
  - max inter-request delay: 59 seconds

Inter-req delay distribution on bored web server



90% of time: < 8sec
99% of time: < 23sec

(And this server is *bored*. On a busy day - pi day - 99% < 4s)

On busy day, with 15sec disk, could have slept for 5 minutes

Real servers

- Idle probably doesn't work at all for a busy-ish server
  - Caveat: Might work really well for departmental/ workgroup/etc. servers
  - Highly load pattern dependent!
- Just using laptop disks - these guys find doesn't work
- More recent work suggests it might. Power gap to modern laptop drives has increased...

Seagate Momentus

- 7200RPM 500GB
- 1.6W active (avg; 2.2W max) 1.4/0.7W idle
- 11ms random read seek time
- Savvio - 2.5" server drive
- 2.9ms random read seek time (avg)
- 8.43W active avg; 5.8W idle
- 5.2x more power, 3.8x faster - but 8x more idle pw

Hard Drives

- Remember seek times:
  - rotational latency (expected 1/2 rotation) + time to move read head & settle
  - 7200RPM drive: 8.3ms per rotation (4.2avg)
  - Move read head: ~3ms avg for 2.5" server drive; ~12ms (!) for 2.5" laptop

Workloads

- Are you CPU bound, seek-bound, or bandwidth-bound?
- Paper examines mostly seek-bound workloads -- proxy caching is a prime example

Variable Speed Drives

- 2007: Hitachi, WD introduce variable speed drives
  - WD: "IntelliSeek" - adapts rotational velocity to seek location (draw picture; for ref, see cute animation on WD website) 5400-7200RPM