


Practical power management for data centres


Dushyanth Narayanan
Microsoft Research, Cambridge



Energy in data centers

- Substantial portion of TCO
 - Power bill, peak power ratings
 - Cooling
 - Carbon footprint
- It's (becoming) a big deal for Microsoft
 - Our own data centers
 - Enterprise customers

Microsoft HP Server Computing Summit 2008 2



Challenge

- Operators/designers have other priorities
 - Performance
 - Cost
 - Availability/reliability
- Need energy to become a top priority
- It is not (yet)
 - Beware of the hype


Microsoft HP Server Computing Summit 2008 3



Keep it simple

- Do not increase complexity (too much)
- Try the obvious thing first
- What can we turn off?
 - Put CPUs in sleep states
 - Spin down disks
 - Power down machines


Microsoft HP Server Computing Summit 2008 4



Go for the 80% solution

- Where does your power go?
 - CPU? Disks? Power supplies? Memory?
- Do you care about peak or mean power?
- Need better power instrumentation
 - Per-machine, per-component
 - Measure under real workloads


Microsoft HP Server Computing Summit 2008 5



Keep it fast

- Do not degrade performance (too much)
- What is the performance impact?
 - CPU sleep state → millisecond(s) delay
 - Disk spin up → seconds
 - Machine power-up → 10s of seconds

Microsoft HP Server Computing Summit 2008 6



Hope the hardware improves

- Flash storage
 - Eventually replace spinning spindles
 - Not there yet (capacity/perf)
 - Can we use hybrid flash/disk systems?
- Picoservers

Microsoft HP Server Computing Summit 2008

7




Roadmap

- Hand-wavy ranting
- Concrete stuff
 - Write off-loading
 - One way to save power in enterprise storage

Microsoft HP Server Computing Summit 2008


8



Why storage?

- Storage is significant
- Especially in an idle system
 - Idle Seagate Cheetah 15K.4: 12 W
 - Idle Intel Xeon dual-core: 24 W

Microsoft HP Server Computing Summit 2008 9



Challenge

- Most of disk's energy just to keep spinning
 - 17 W peak, 12 W idle, 2.6 W standby
- Flash still too expensive
 - Cannot replace disks by flash
- So: need to spin down disks when idle

Microsoft HP Server Computing Summit 2008 10

Microsoft
Research

Intuition

- Real workloads have
 - Diurnal, weekly patterns
 - Idle periods
 - Write-only periods
 - Reads absorbed by main memory caches
- We should exploit these
 - Convert write-only to idle
 - Spin down when idle

Microsoft HP Server Computing Summit 2008 11


Microsoft
Research

Small/medium enterprise DC

- 10s to 100s of disks
 - Not MSN search
- Heterogenous servers
 - File system, DBMS, etc
- RAID volumes
- High-end disks

The diagram shows a server rack with three server units. The top unit is labeled 'FS1' and is connected to two RAID volumes, 'Vol 0' and 'Vol 1'. The middle unit is labeled 'FS2' and is connected to three RAID volumes, 'Vol 0', 'Vol 1', and 'Vol 2'. The bottom unit is labeled 'DBMS' and is connected to two RAID volumes, 'Vol 0' and 'Vol 1'. Each RAID volume is represented by a group of four disk icons.


Microsoft HP Server Computing Summit 2008 12



Design principles

- Incremental deployment
 - Don't rearchitect the storage
 - Keep existing servers, volumes, etc.
 - Work with current, disk-based storage
 - Flash more expensive/GB for at least 5-10 years
 - If system has some flash, then use it
- Assume fast network
 - 1 Gbps+


Microsoft HP Server Computing Summit 2008 13



Write off-loading

- Spin down idle volumes
- Offload writes when spun down
 - To idle / lightly loaded volumes
 - Reclaim data lazily on spin up
 - Maintain consistency, failure resilience
- Spin up on read miss
 - Large penalty, but should be rare

Microsoft HP Server Computing Summit 2008 14




Roadmap

- Motivation
- Traces
- Write off-loading
- Evaluation

Microsoft HP Server Computing Summit 2008

15



How much idle time is there?

- Is there enough to justify spinning down?
 - Previous work claims not
 - Based on TPC benchmarks, cello traces
 - What about real enterprise workloads?
 - Traced servers in our DC for one week

Microsoft HP Server Computing Summit 2008

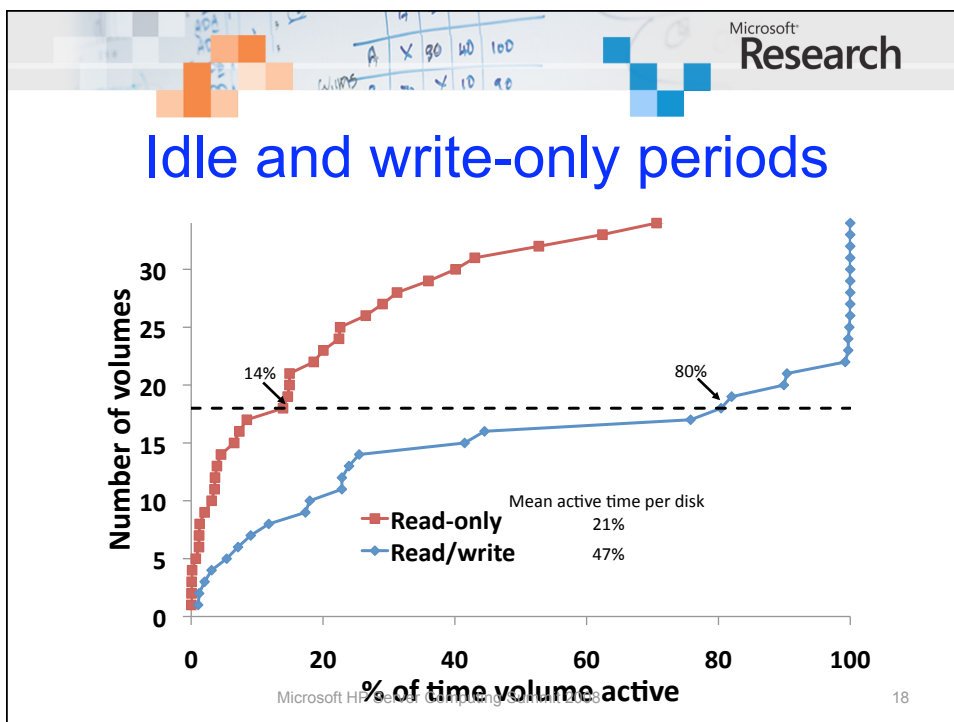
16


Microsoft
Research

MSRC data center traces

- Traced 13 core servers for 1 week
 - File servers, DBMS, web server, web cache, ...
 - 36 volumes, 179 disks
 - Per-volume, per-request tracing
 - Block-level, below buffer cache
- Typical of small/medium enterprise DC
 - Serves one building, ~100 users
 - Captures daily/weekly usage patterns

Microsoft HP Server Computing Summit 2008 17






Roadmap

- Motivation
- Traces
- Write off-loading
- Preliminary results

Microsoft HP Server Computing Summit 2008 19



Write off-loading: managers

- One manager per volume
 - Intercepts all block-level requests
 - Spins volume up/down
- Off-loads writes when spun down
 - Probes logger view to find least-loaded logger
- Spins up on read miss
 - Reclaims off-loaded data lazily

Microsoft HP Server Computing Summit 2008 20

Microsoft
Research

Write off-loading: loggers

- Reliable, write-optimized, short-term store
 - Circular log structure
- Uses a small amount of storage
 - Unused space at end of volume, flash device
- Stores data off-loaded by managers
 - Includes version, manager ID, LBN range
 - Until reclaimed by manager
 - Not meant for long-term storage


Microsoft HP Server Computing Summit 2008 21

Microsoft
Research

Off-load life cycle

The diagram illustrates the off-load life cycle. On the left, a dashed box represents a local storage environment containing two cylinders labeled 'v2'. A blue arrow labeled 'Write' points to the top cylinder, and a green arrow labeled 'Spindown' points to the bottom cylinder. A green arrow points from this local storage to a dashed box on the right labeled 'v1', which contains two cylinders. A green arrow points from 'v1' to a dashed box at the bottom containing three cylinders. A green arrow points from the bottom box back to 'v1', and another green arrow points from 'v1' back to the local storage 'v2'. The text 'I/O Write' is positioned above the arrow connecting the local storage to 'v1'.


Microsoft HP Server Computing Summit 2008 22



Consistency and durability

- Read/write consistency
 - manager keeps in-memory map of off-loads
 - always knows where latest version is
- Durability
 - Writes only acked after data hits the disk
- *Same guarantees as existing volumes*
 - Transparent to applications


Microsoft HP Server Computing Summit 2008 23



Recovery: transient failures

- Loggers can recover locally
 - Scan the log
- Managers recover from logger view
 - Logger view is persisted locally
 - Recovery: fetch metadata from all loggers
 - On clean shutdown, persist metadata locally
 - Manager recovers without network communication


Microsoft HP Server Computing Summit 2008 24



Recovery: disk failures

- Data on original volume: same as before
 - Typically RAID-1 / RAID-5
 - Can recover from one failure
- What about off-loaded data?
 - Ensure logger redundancy \geq manager
 - k-way logging for additional redundancy


Microsoft HP Server Computing Summit 2008 25



Roadmap

- Motivation
- Traces
- Write off-loading
- Experimental results


Microsoft HP Server Computing Summit 2008 26



Testbed

- 4 rack-mounted servers
 - 1 Gbps network
 - Seagate Cheetah 15k RPM disks
- Single process per testbed server
 - Trace replay app + managers + loggers
 - In-process communication on each server
 - UDP+TCP between servers


Microsoft HP Server Computing Summit 2008 27



Workload

- Open loop trace replay
- Traced volumes larger than testbed
 - Divided traced servers into 3 “racks”
 - Combined in post-processing
- 1 week too long for real-time replay
 - Chose best and worst days for off-load
 - Days with the most and least write-only time


Microsoft HP Server Computing Summit 2008 28



Configurations

- Baseline
- Vanilla spin down (no off-load)
- Machine-level off-load
 - Off-load to any logger within same machine
- Rack-level off-load
 - Off-load to any logger in the rack

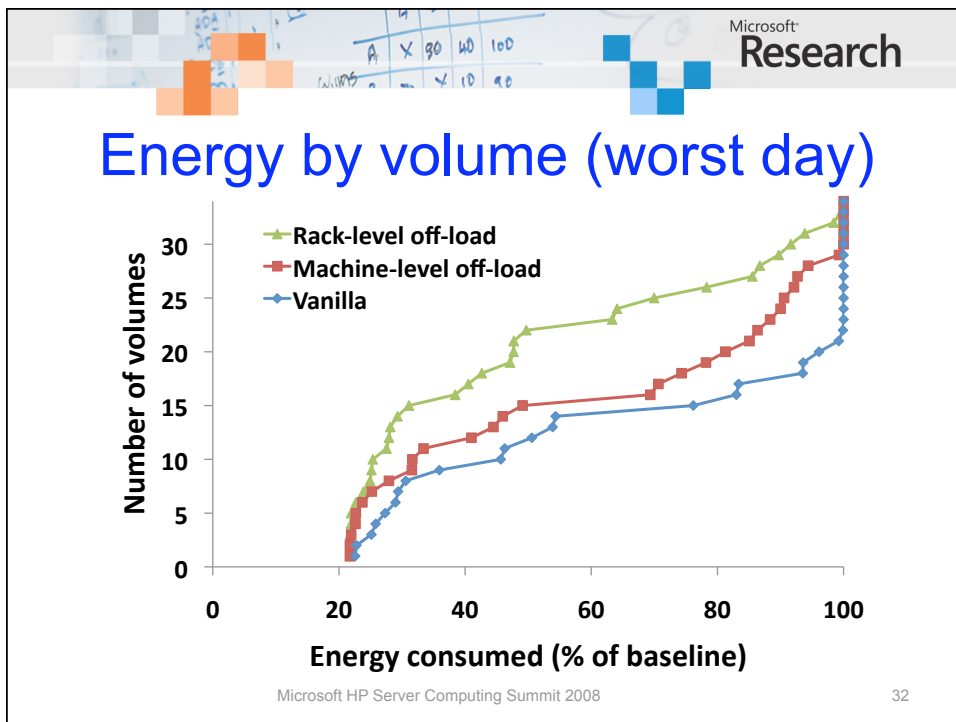
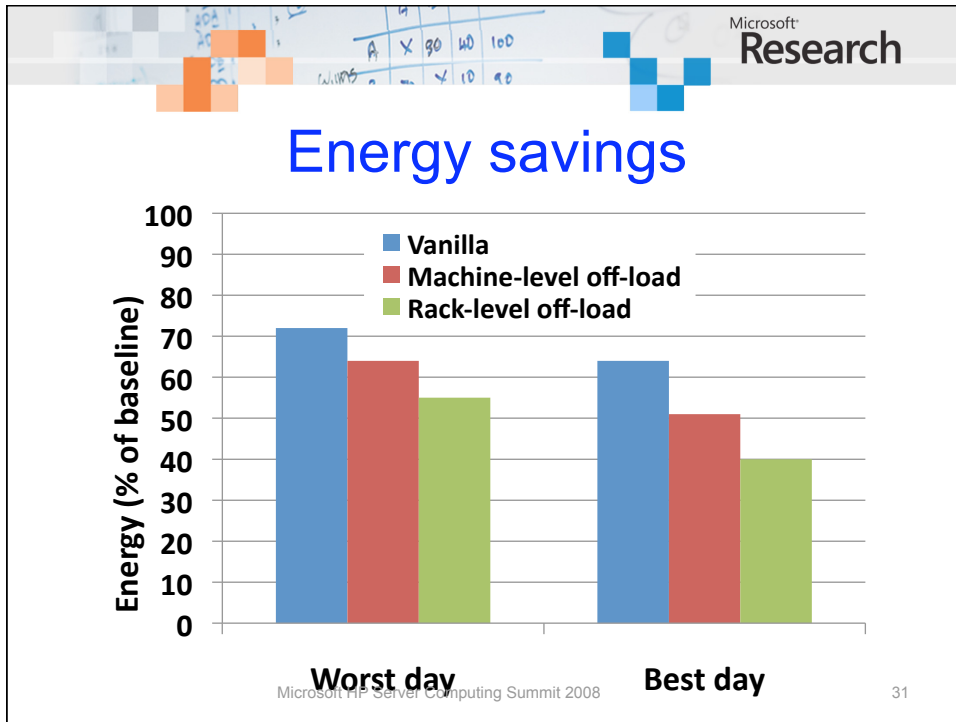
Microsoft HP Server Computing Summit 2008 29

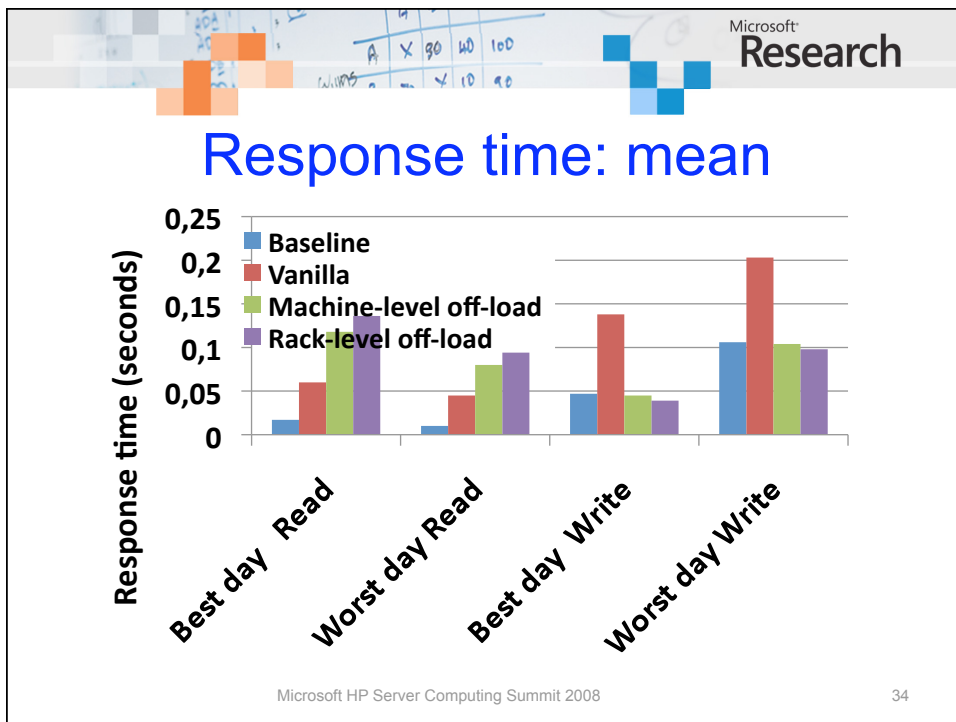
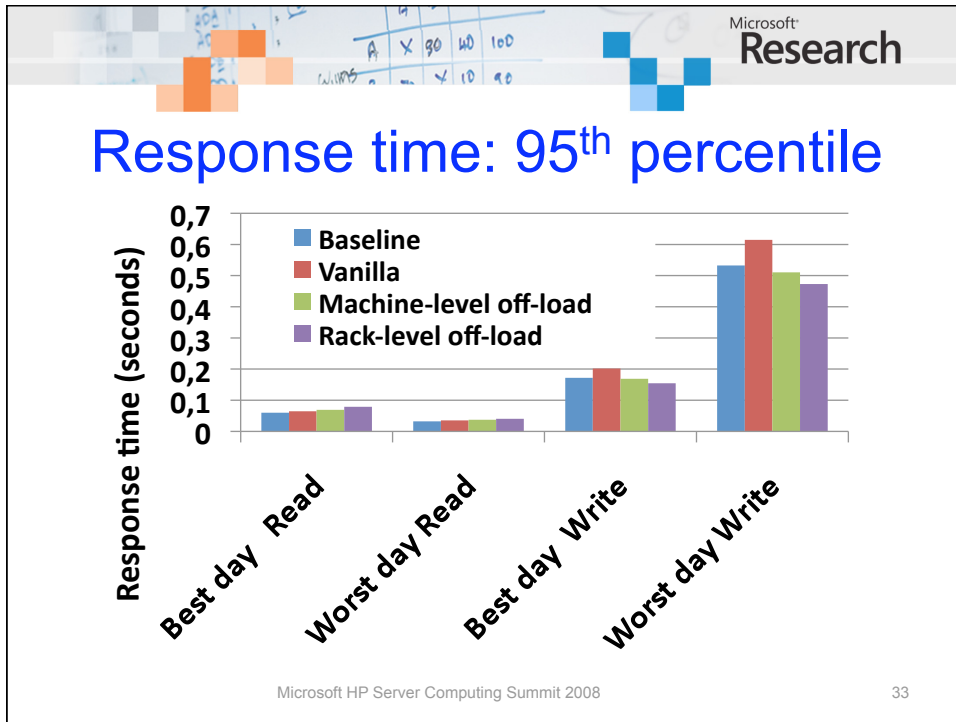



Storage configuration

- 1 manager + 1 logger per volume
 - For off-load configurations
 - Logger uses 4 GB partition at end of volume
- Spin up/down emulated in s/w
 - Our RAID h/w does not support spin-down
 - Parameters from Seagate docs
 - 12 W spun up, 2.6 W spun down
 - Spin up delay is 10—15s, energy penalty is 20 J
 - Compared to keeping the spindle spinning always

Microsoft HP Server Computing Summit 2008 30







Conclusion


- Need to save energy in DC storage
- Enterprise workloads have idle periods
 - Analysis of 1-week, 36-volume trace
- Spinning disks down is worthwhile
 - Large but rare delay on spin up
- Write off-loading: write-only → idle
 - Increases energy savings of spin-down

Microsoft HP Server Computing Summit 2008 35



Questions?


Microsoft HP Server Computing Summit 2008 36



Related Work

- PDC
 - ↓ Periodic reconfiguration/data movement
 - ↓ Big change to current architectures
- Hibernator
 - ↑ Save energy without spinning down
 - ↓ Requires multi-speed disks
- MAID
 - Need massive scale

Microsoft HP Server Computing Summit 2008 37



Just buy fewer disks?

- Fewer spindles → less energy, but
 - Need spindles for peak performance
 - A mostly-idle workload can still have high peaks
 - Need disks for capacity
 - High-performance disks have lower capacities
 - Managers add disks incrementally to grow capacity
 - Performance isolation
 - Cannot simply consolidate all workloads

Microsoft HP Server Computing Summit 2008 38

