OPTIMIZING THE ENSEMBLE

by: N. Tolia et. al.

Motivation

* power is a limiting factor in data centers
* for a variety of reasons, data centers are over-provisioned (utilization < 100%)
* CPU power control advances mean other system components (including cooling) now dominate
* ideally we would like energy proportional systems: 0 Watts at 0% utilization, Max Watts at 100% util.

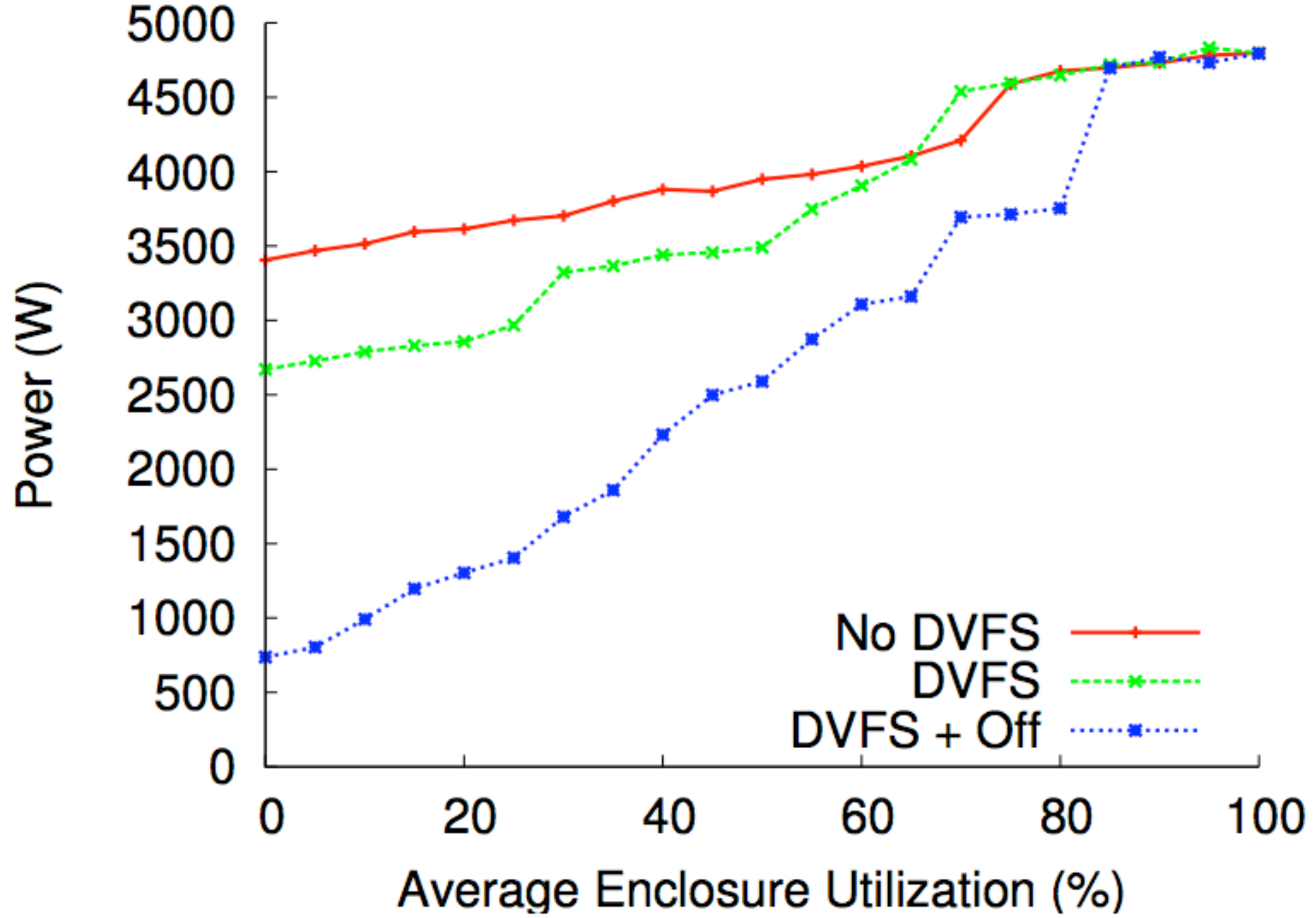
Experimental Setup

* blade enclosure with 16 blades and 10 fans
* each blade has 2 dual-core AMD CPUs and 16 GB of RAM
* system is not energy proportional (high idle power)
* Xen VMs + SAN (fibre channel to a consolidated storage server)
* workload: 'gamut' generates target load levels

Blade Energy Proportionality

* No DVFS - no power saving techniques
* DFVS - scaling in reaction to load (like Linux OnDemand governor)
* DFVS + Off - also migrate VMs (with a CPU and memory utilization constraint) and power off servers
* all 64 VMs experience the same load level

Proportionality Achieved

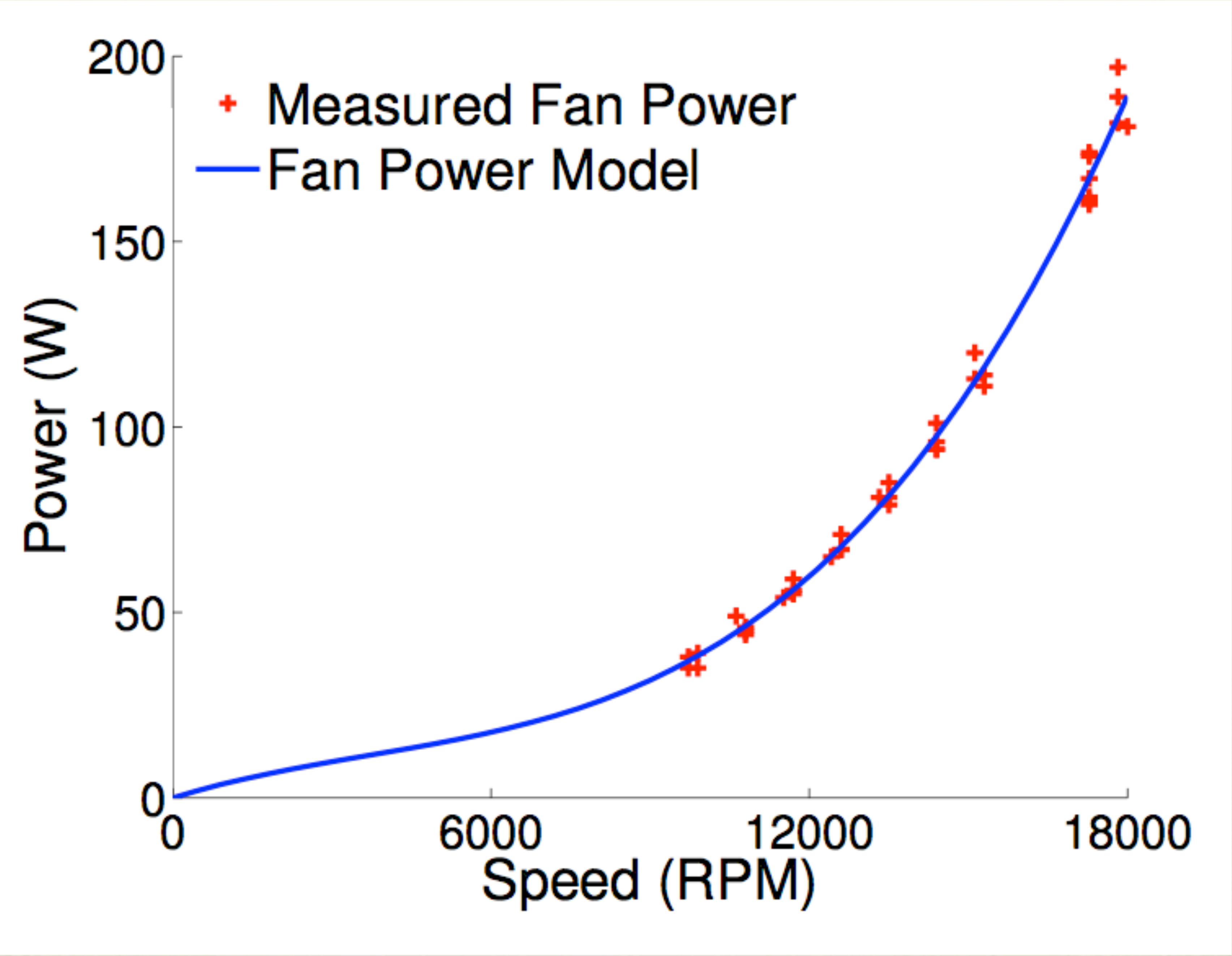


Each result presented above is an average of approximately 90 readings over a 15 minute interval.

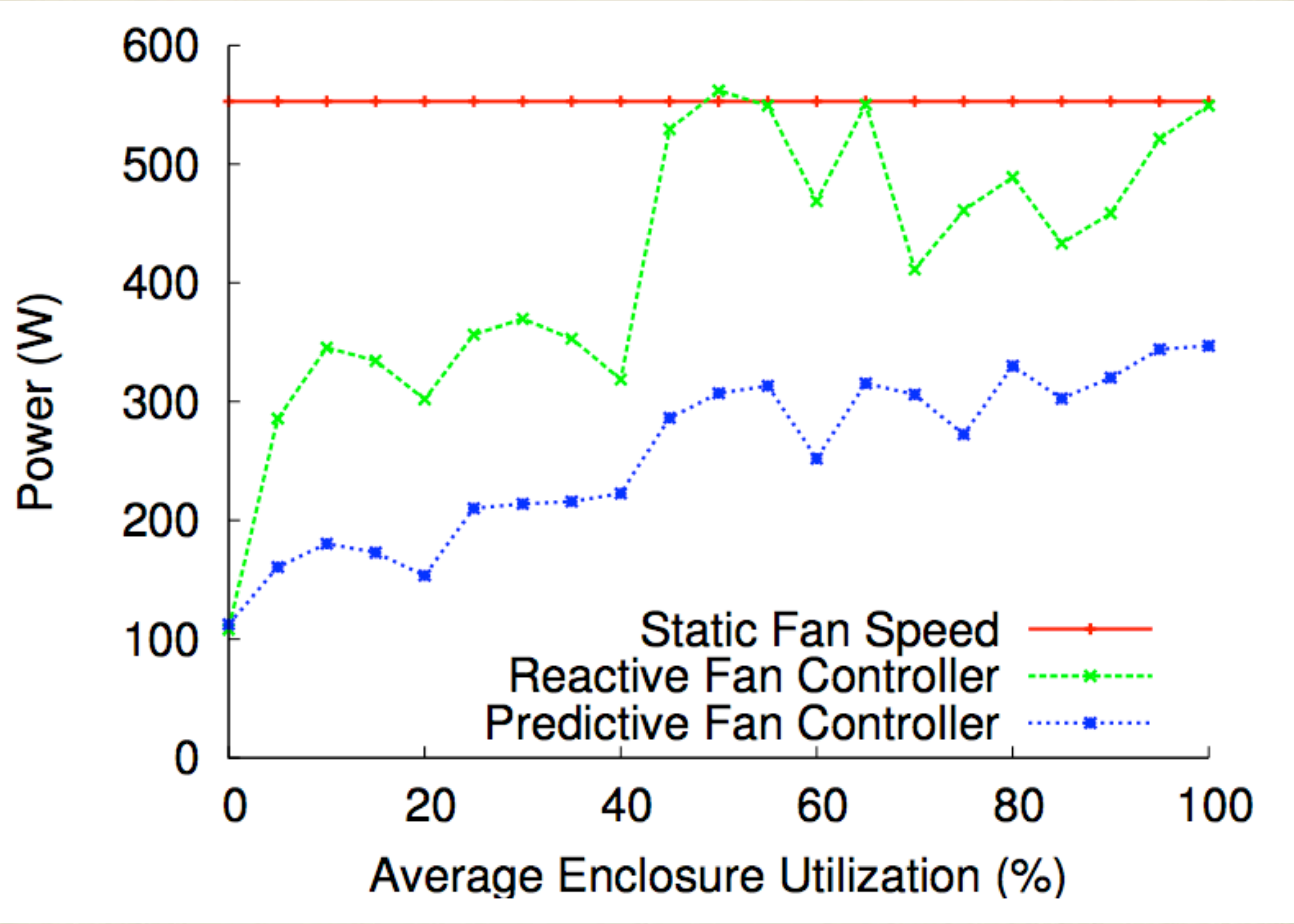
Cooling Energy Proportionality

* Server fans can consume 10-25% of server power
* 10 fans cool 16 blades in enclosure
* always on and thermally reactive fan control policies are not proportional
* predictive policy uses load information to adjust cooling for specific blades

Cubic Fan Power



Proportional Cooling?



Summary

* Managing non-energy proportional systems in aggregate can lead to more proportional behavior
* speed control and on-off are needed together to do so

ENERGY PROPORTIONALITY FOR STORAGE

by: Jorge Guerra et. al

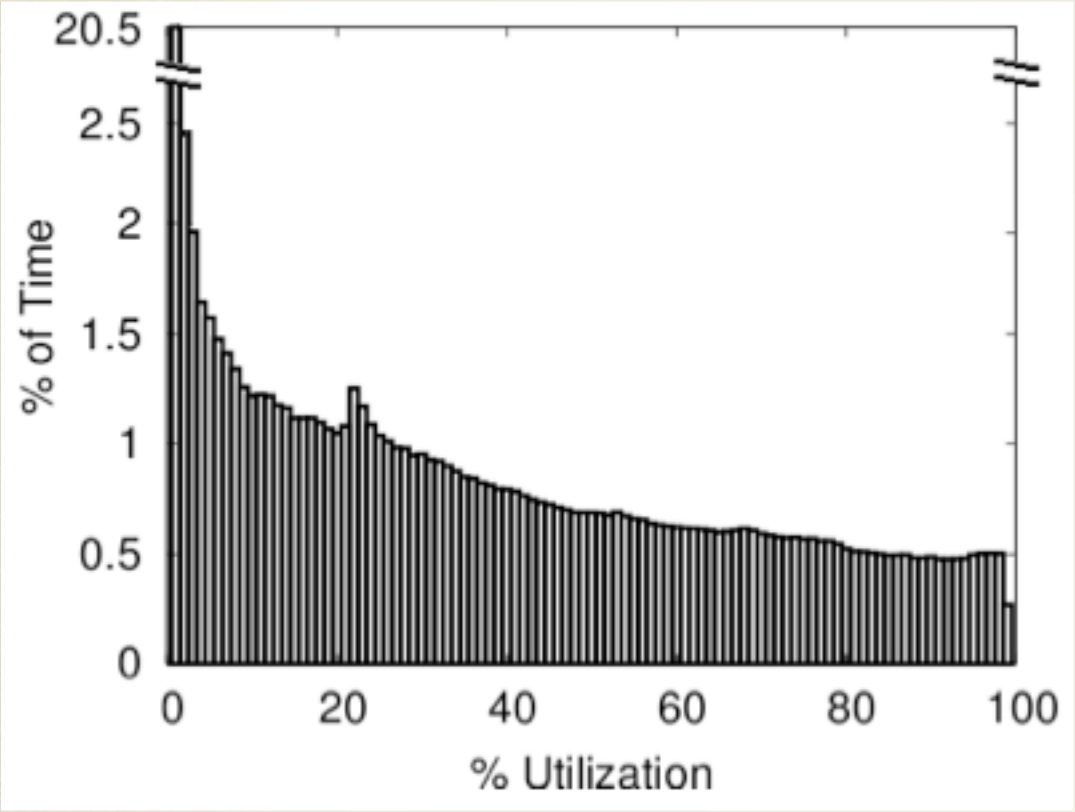
Motivation

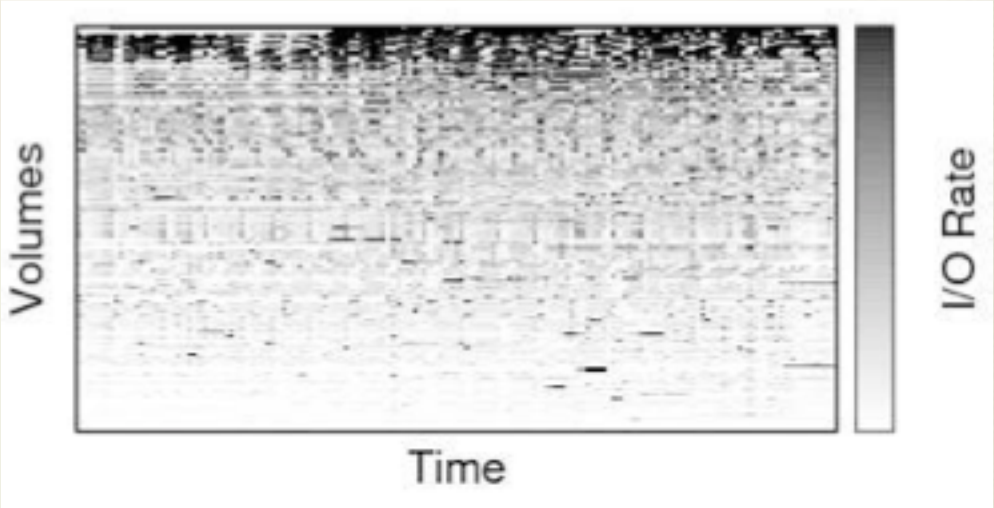
- storage consumes 37-40% of data center IT power
- in the future number of drives (@ 15-20 W) acquired will outstrip number of CPUs (@3-20W) acquired:
  - slow capacity improvements
  - move to 2.5 inch drives (more J/GB)
  - performance lags capacity (short stroking)
- energy efficiency isn't enough, we need energy proportional storage

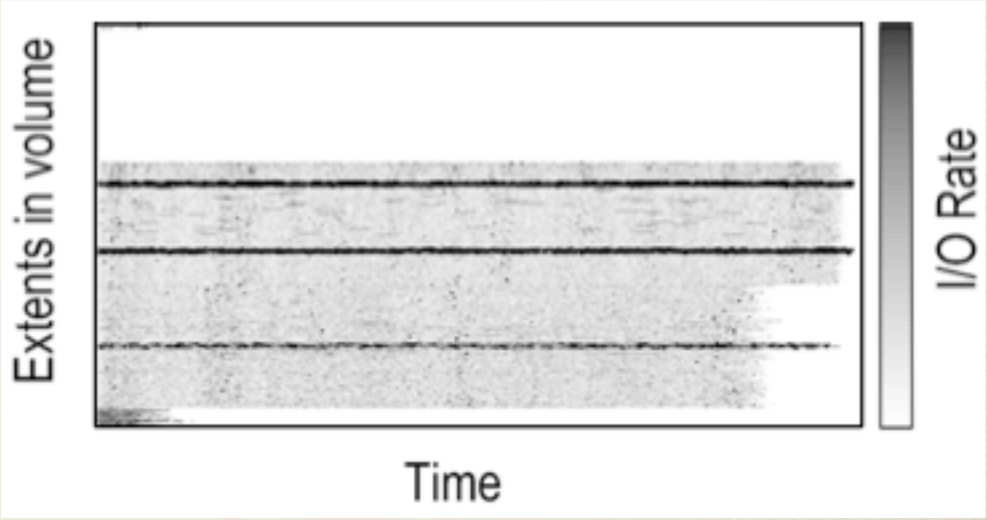
Two Optimization Scenarios

* performance matters most
  * energy use should vary with performance requirement
* energy matters most
  * maximum performance given constraint
  * this is becoming the more relevant scenario

Exploitable Variation Exists



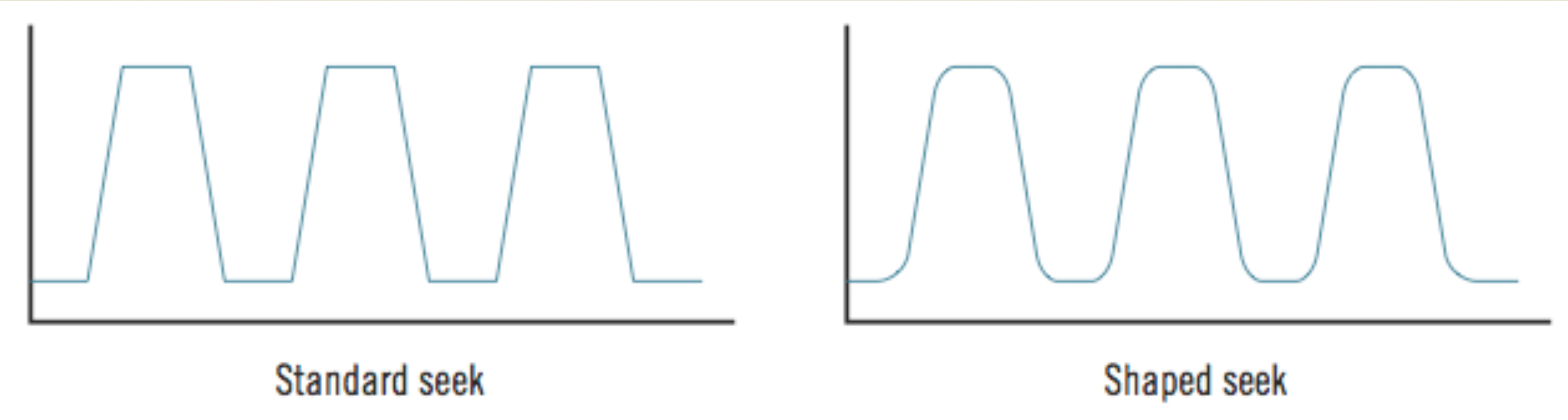


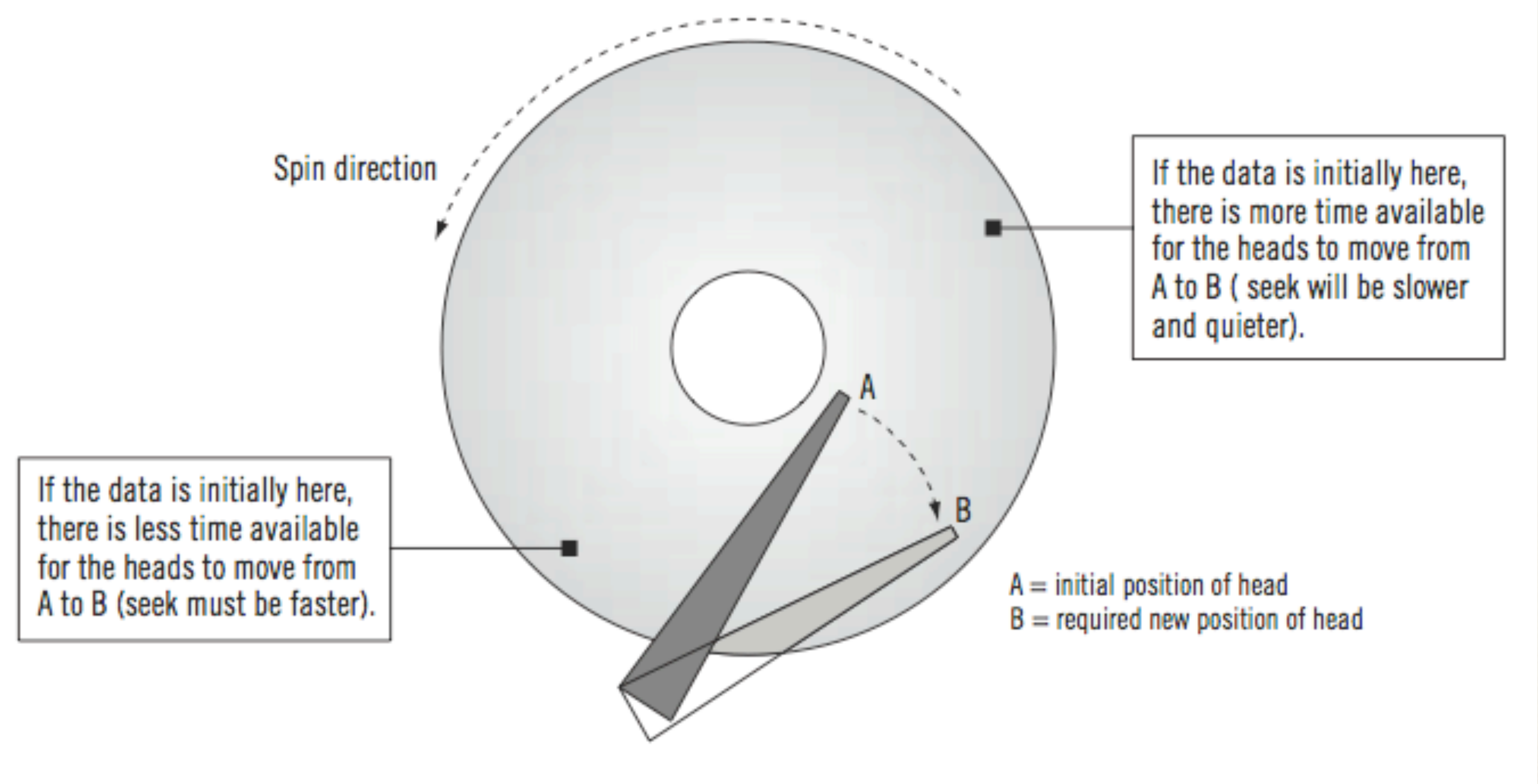


Using Disk Power Modes

* nothing like DFVS exists for disks (DRPM notwithstanding) so what can we do?
* Opportunistic Spindown: stop spinning platters after a given idle period (rent-to-buy)
* Workload Shaping: batch I/O requests to produce longer idle periods (prefetching, read-ahead, app-level)
* Changing Seek Speed: alter velocity and/or acceleration of seeks to reduce noise (also power). JIT seeks.

Shaped and JIT Seeks





Placement and Migration Techniques

- Consolidation: colocation and avoiding short stroking
- Tiering/Migration: Enterprise and SATA drives, SSDs
  - putting the 4% most popular extents on SSD and the remaining on SATA can save 75% power of using all Enterprise disks for the same cost
- Dedup/Compression: store less data

Placement and Power Modes

- Spindown + Write Offloading: don't wake up disks for writes (writes must be cached persistently)
  - a kind of workload shaping
- Spindown + MAID/PDC: reorganize popular data onto a subset of disks, hope other disks are mostly idle

Requirements

* high sensitivity to peak Response Time and average RT
  * critical business apps, transactional databases
* low peak RT sensitivity, high average RT sensistivity
  * multimedia streaming, file storage
* low peak and average RT sensitivity
  * archival/backup and SarbOx compliance

Time and Space Granularity

Technique	App Category	Time-scale	Granularity	Potential to alter performance
Consolidation	1,2,3	hours	coarse	Can lengthen response times
Tiering/migration	1,2,3	minutes-hours	coarse	Can lengthen response times
Write off-loading	2,3	milliseconds	coarse	Adds background process that can impact application
Adaptive seek speeds	1,2,3	milliseconds	fine	Can lengthen response times
Workload shaping	2,3	seconds	fine	Can lengthen response times
Opportunistic spindown	2,3	seconds	fine	Delays due to spinup
Spindown/MAID	3	10's of seconds	medium	Delays due to spinup
Dedup/compression	2,3	n/a	n/a	Delays in accessing data due to assembling from repository or decompression

Table 2: Volume categorization for the financial data center workload. Key: *H*: high load, *L*: low load, *P*: peaks in load, *V*, V_X: variable load (V_1=lowest, V_4=highest I/O rate).

Category	H	L	P	V	V_1	V_2	V_3	V_4
% Vol.	10	5	13	72	51	6	4	11

Table 3: Framework for mapping storage application performance requirements and workload characteristics to energy saving techniques. Techniques: C: *Consolidation*, T: *Tiering/Migration*, S: *Opportunistic Spin-down/MAID*, W: *Write Off-loading*, A: *Adaptive Seek Speeds*, H: *Workload Shaping*, D: *Dedup/Compression*.

Sensitivity to Avg. Resp. Time	Sensitivity to Peak Resp. Time	Stability of Workload	Techniques						
			C	T	S	W	A	H	D
Yes	Yes	No							
		Yes	✓	✓					
	No	No			✓	✓			
		Yes	✓	✓	✓	✓			
No	No	No			✓	✓	✓	✓	✓
		Yes	✓	✓	✓	✓	✓	✓	✓

Key ✓ : Applicable.

Conclusion

* Real world I/O workload analysis is encouraging for our ability to apply power saving techniques (40% savings for energy-proportional volume trace)
* if we have workload stability or can tolerate occasional delays, power saving techniques exist
* if we can tolerate an increase in average response time a wide variety of techniques are at our disposal