Correcting vibration-induced performance degradation in enterprise servers

Christine S. Chan[†], Boxiang Pan[‡], Kenny Gross[§], Kalyan Vaidyanathan[§], Tajana Šimunić Rosing[†,‡]

[†]*Dept. of Electrical and Computer Engineering,* [‡]*Dept. of Computer Science and Engineering*
University of California, San Diego, CA
[§]*Oracle Physical Sciences Research Center, San Diego, CA*
{csc019, bopan, tajana}@ucsd.edu, {kenny.gross, kaylan.vaidyanathan}@oracle.com

ABSTRACT

Server fan subsystems are power-hungry and generate vibrations, degrading the performance of data-intensive workloads and inflating the uptime electric bills of cost-sensitive datacenters. In this paper, we demonstrate a systematic server measurement methodology to isolate different types of vibrations and quantify their effect on hard disks. We introduce a thermal and cooling management policy that mitigates vibrational effects workload scheduling and fan control, and study the benefits of a hybrid storage array with solid-state drives (SSDs) that are impervious to vibrations. We achieve performance improvements of up to 73% and energy savings of up to 76% over the state of the art, while meeting thermal constraints and improving the system's resilience to both internal and external vibrations.

1. INTRODUCTION

Modern usage of computing has led to ballooning file sizes and demand for high performance in real-time services. As developments in processing power have advanced far ahead of storage and communication, many services are now I/O bound. Datacenters support interactive web applications, OLTP database operations, cloud services and market trading. Owners such as Amazon EC2, Facebook, Google and stock exchanges all have mission-critical workloads and need to guarantee quality of service (QoS) to their customers and fulfill the terms of their service level agreements (SLA).

Ever shrinking transistor feature sizes lead to increased power densities and higher temperatures. To protect hardware components and maintain runtime performance, datacenters have high-powered HVAC systems in the buildings and fan subsystems in server chassis to maintain a thermal set point. However, high fan speeds have a high cost directly and indirectly. Their cubic power profile can account for up to 28-51% of total server energy consumption in enterprise applications [1]. A lesser known cost is the performance degradation of hard disks caused directly by vibrations generated inside a server chassis and transmitted through racks [2]. Seagate identifies rotational vibration from disk actuation and external forces as an area of concern for disk data transfer speeds [3]. The American Society of Heating, Refrigeration, and Air-Conditioning Engineers (ASHRAE) recommends that datacenters save on IT operational costs by turning down computer room air conditioning units (CRACs)[5], but these IT energy savings may be offset by the high motor power of fans and elevated vibrations degrading IO performance.

In today's hard disk drives, each disk platter spins at up to 15,000 rotations per minute (RPM), while the read/write head, 7 nanometers away, is targeting tracks 20 nanometers wide. Because of this areal density, any vibrational disturbance to the storage array causes one or multiple sequences of read-retries and write-retries. Reissuing these access requests lowers disk transfer rates, which delays application performance and threatens the QoS guarantees of time-critical service jobs. This delay can also be quantified in the extra energy cost of powering server components to support a prolonged execution time for any given customer workload.

Current vibration-canceling mechanisms only target regular external vibrations and cannot react quickly enough to the changes introduced by variable fan speeds. Enterprise hard disks are manufactured for density – there is no space left in the server chassis to accommodate damping materials. Vibration-induced I/O performance degradation is very difficult to diagnose in deployment. The frequency band of structural resonances inside servers or the racks that hold servers can be very narrow (sometimes down to 2Hz). A variable-speed fan may intersect the structural resonance and amplify vibration amplitudes, raising latencies in database transactions suddenly, and leading to time-outs hangs at the customer level. If such a server is returned under a service contract to the vendor to be tested in a repair center with a different resonance frequency, the degradation is difficult or impossible to reproduce. Thus, the server might be reported as NTF – "no trouble found" – and have to be replaced. This workaround extends root-cause analysis timelines, incurs high warranty costs for enterprise IT systems and does nothing to resolve the underlying problem. Runtime solutions are required to combat such environmental variability, orthogonal to innovations in better hardware enclosures.

Today's thermal management and workload scheduling policies manage core scheduling, memory page scheduling, and fan speed control separately, leading to temperature hotspots on-chip and energy inefficient solutions. Our novel thermal management policy takes into account the thermal interactions between CPUs and memory modules, and mitigates the effect of fan-induced vibrations on disk performance. By introducing performance awareness in the form of characterization curves between disk throughput and fan speeds, the controller can make more intelligent choices in setting fan speeds based on sensitivity of workloads to disk performance as well as temperatures. Indeed, as SSDs get introduced into enterprise datacenters, it is most often as a "cache" layer in front of conventional spinning disks. We leverage a hybrid SSD storage set up to improve system resilience

GreenMetrics '13, June 17, 2013, Pittsburgh, PA, USA.

to external vibrations, in addition to the software mitigation of internal vibrations through the fan controller. We show the benefits of our solution based on metrics important to datacenters: service performance and energy savings.

2. RELATED WORK

In general, it is acknowledged that vibrations and shock are non-ideal for the operation of a hard disk drive. Most systems have built-in protection from rare shock events. Some enterprise drives (e.g. Seagate Constellation ES.3) advertise compensation or tolerance for vibrations, but they are only rated for predictable, consistent frequencies and amplitudes. Their solutions integrating SCSI, cache architectural design and plurality in multi-drive systems focus on reducing drive-emitted instead of externally produced vibrations, and as such do not resolve issues we observed on our measurements.

At the heart of this problem is inefficient cooling and thermal management due to multiple independent controllers. The authors of [6] introduce a thermal model and MIMO fan control, mentioning fan-related resonance in passing, but do not investigate the source of such vibrations. The fan controller yields good power efficiency but ignores total runtime energy consumption performance. Dynamic voltage and frequency scaling (DVFS) reduces temperatures at the expense of proportional performance overhead [7]. Several new schemes integrate power and thermal management [8][9] with a focus on processor and memory energy efficiency [10]. These publications and even studies of disk-heavy database query performance [11] focus only on quantifying CPU performance without investigating the effects of disk performance or the energy costs. The authors of [2] present a thermal model of enterprise servers and introduce the relationship between server fans and hard drive throughput, demonstrating some possible energy gains of reducing fan vibrations. However, they assume a simple relationship for the fan speed vs. disk performance, ignoring external variations and variability due to different disk types. Also, their solution requires tight application-level integration into low-level thermal management in conjunction with intrusive I/O monitoring, which is unrealistic for today's systems.

In terms of per-gigabyte costs, it is still economically infeasible for solid-state drives (SSDs) to replace legacy spinning storage in cost-sensitive datacenters [4]. However, database software providers use SSDs as an extra layer in the memory hierarchy to increase the size of the main memory buffer [17], serving as temporary caching instead of permanent storage. This new trend may reduce a system's dependence on spinning hard drives and thus lessen the effects of fan-induced vibrations on the disks.

In the light of these observations and upcoming trends, we present the following contributions:

- In contrast to [2], we present detailed measurements showing that disk sensitivity to vibrations is not a simple function of fan speeds, but the dependence on vibrational amplitudes and frequencies are much more complex.
- We present a thermal and cooling management policy that dynamically responds to internally generated vibrations, and improves resilience to external vibrations through SSD-based caching.

Section III describes our measurement methodology. Several state of the art and new power and thermal management policies are described in Section IV, and are evaluated in Section V.

3. METHODOLOGY

In this section we present a methodology for characterizing any server disk's response to vibrations and the resulting effect on the overall server performance. We introduce our setup for isolating and measuring the different vibrations hard drives will experience in a datacenter.

3.1 Instrumentation

The test platform is a state-of-the-art single-socket server at 40nm with 8 multi-threaded cores (8 threads each), two memory sockets on either side of the processor, 6 fan modules, and 8 disk drive slots. The boot drive is reserved for the measurement infrastructure and workload scripts (lower left drive in Figure 1). The other slots are loaded with a broad range of disk models as described in Table 1, including sensitive SATA drives (commodity drives most preferred in datacenters), more resilient SAS2 drives, and an SSD impervious to vibrations. In a real system, the bays would be filled with the same model. The server consumes between 330W-600W depending on utilization, while the socket itself has maximum thermal design power (TDP) of 240W. The fans provide a maximum airflow of 145 cubic feet per minute (cfm) at 180W. Room temperature is maintained through an HVAC system at 25°C, and chassis internal ambient temperature is 30°C.

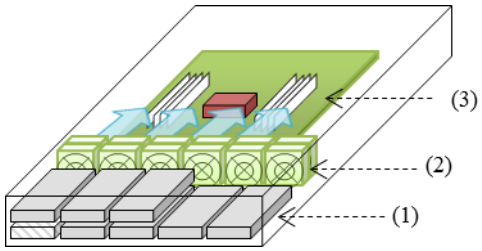

Figure 1. Server organization with (1) hard disks and (2) fan assembly directing airflow towards (3) the motherboard.

We use tri-axial accelerometers to measure vibrations in server racks within an operative datacenter, then reproduce the environment with an Unholtz-Dickie model K170 electrodynamic programmable vibrational table. Fans are controlled via the Intelligent Platform Management Interface (IPMI). Fan speeds are described through pulse width modulation (PWM), where the "pulse width" refers to the duty cycle of a digital signal. (This

Table 1 Disk drive models measured

Model	Type	Spin speed (RPM)	Capacity (GB)	Measured write speed (MB/s)	Abbreviation
Seagate Savvio 10K.3 ST930003S	SAS	10000	300GB	72.2	SEA SAS2 A
Seagate Savvio 10K.3 ST930003S	SAS	10000	300GB	70.6	SEA SAS2 B
Hitachi Ultrastar C10K600	SAS	10000	600GB	81.6	HIT SAS
Fujitsu MHY2200BS	SATA	5400	200GB	31.2	FUJ SATA
Hitachi Travelstar E5K500	SATA	5400	500GB	37.0	HIT SATA
Intel 710 SSDSA2BZ300G3	SSD	-	300GB	206.0	INT SSD

electrical "pulse" does not contribute to mechanical vibrations.) The rotational fan speed and the resulting air flow are generally linear with the PWM setting, except at either end of the spectrum. Cooling efficiency decreases at high fan speeds, since convective resistance of the packaging (ability to dissipate heat) is inversely proportional to fan speeds [12].

3.2 Measurements

We run a parametric characterization suite of experiments to measure the vibrational sensitivity of a diverse set of disks. In this section, we disable the buffer cache that would have hidden disk access latency from the user. We run a pure IO generator which issues random writes to the disk, utilizing 100% of the I/O bus bandwidth to expose and isolate the effect that vibrations have on disk throughput. The impact of disk throughput degradation on the overall performance of the realistic database benchmarks varies depending on the behavior of each workload. These are evaluated with re-enabled buffer caches in Section 5.

Fan sweep test: For this test, the server is bolted to the stationary shake table. To study the effect of internal vibrations, we sweep through the range of possible fan speeds while monitoring the average write throughput with the pure IO generator. We step through fan speeds from 100% to 0% PWM at 10% step sizes to obtain stable results and avoid inconsistencies caused by quantization errors. With each change in stimuli, the disk drive throughputs take 20 seconds to respond. In our experience, the processor shuts down within 10 seconds of turning off the fans, while self-reporting on-die temperatures up to 91°C immediately before crashing. Consequently, it is challenging to accurately measure system characteristics in fine-grained steps at low fan speeds. Figure 2 shows the average degradation of write throughput on fan speeds, normalized to the maximum throughput measured on each disk. There are no observable vibrational effects below 50% PWM. SATA drives show the most throughput degradation, down to 35% and 12% of their maximum value. The SAS2 drives show degradation only at the maximum fan setting – HIT SAS loses about 2% of its throughput. SSDs show no performance response to fan speeds - they consistently give a throughput of 206MB/s regardless of environment input.

Next, we study the effect of external vibrations by reproducing the range of frequencies and amplitudes obtained from a real operative datacenter. The vibrations are generated on the shake table while fan speeds are set to 50% PWM.

Amplitude test with random frequencies: We ran experiments on the disks under profiles that cover a different collection of frequencies (200-800Hz in Figure 3 and 200-2000Hz in Figure 4), using the root-mean-square (RMS) of their component signals to

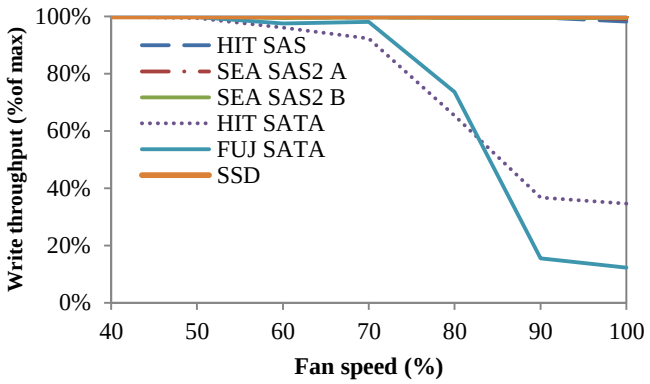

Figure 2 Throughput dependence on fan speeds

fairly compare their response to the total combined signal strength. Although lower throughputs generally follow higher amplitudes, the sensitivity curve varies across hard drives and across frequency profiles. Of the two SATA drives spinning at the same speed (5400 RPM), FUJ SATA performs better than HIT SATA for grms >0.2. At grms =0.63, HIT SATA writes at 8.6 MB/s in the first profile and 3 MB/s in the second. SAS drives are more resilient, but they start showing signs of performance degradation around grms=1.27. The largest drop among the SAS drives is 10.5% on HIT SAS and the largest drop among the SATA drives when HIT SATA stalls at 0 MB/s at grms=1.8. Again, SSDs show no performance response to vibrations and their throughputs are excluded from graphs for clarity.

Frequency test with fixed amplitude: This experiment characterizes the hard disk response to external vibrations of varying frequencies. From on-site measurements at datacenters and observing Figure 3 and Figure 4, we fixed the amplitude of vibrations at 0.17g, where drives performed well in general, but had the potential to experience throughput degradation. We sweep through frequencies between 20 to 2000Hz and monitor the change in disk throughput (Figure 5). The response to different frequencies is irregular and there is neither a distinct "zone" of performance degradation, nor any obvious ratio between the frequency value or write throughput. Certain frequencies that cause performance degradation have a very narrow band. Even though more obvious degradation is seen at higher frequencies, there are narrow bands where disk performance returns close to its ideal. SATA drive throughput drops to 0MB/s at various points, while SAS drives fluctuate by 1-2%.

With these experiments, we have characterized the relationship between hard disk performance and vibrations internal and external to the server. In the next section, we will discuss several

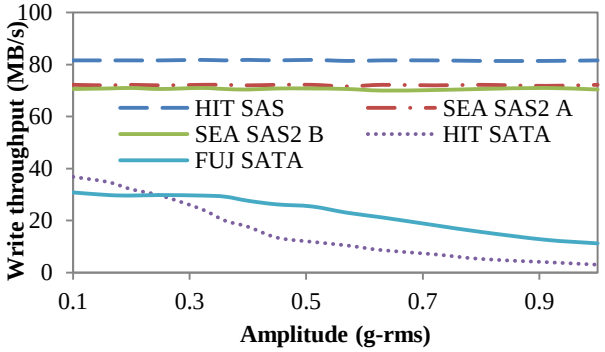

Figure 3. Throughput dependence on amplitude of random vibrations with a frequency profile ranging from 20-800Hz

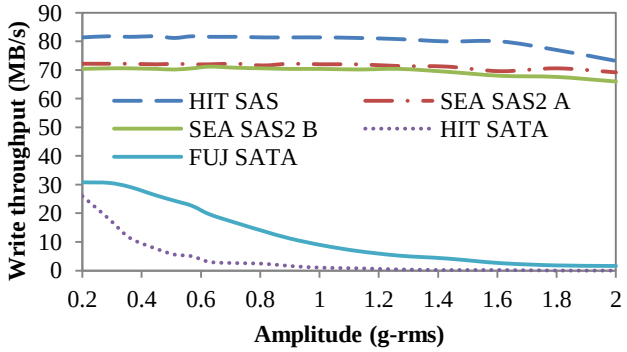

Figure 4. Throughput dependence on amplitude of random vibrations with a frequency profile ranging from 20-2000Hz

dynamic management schemes whose goal is to mitigate hard disk performance degradation.

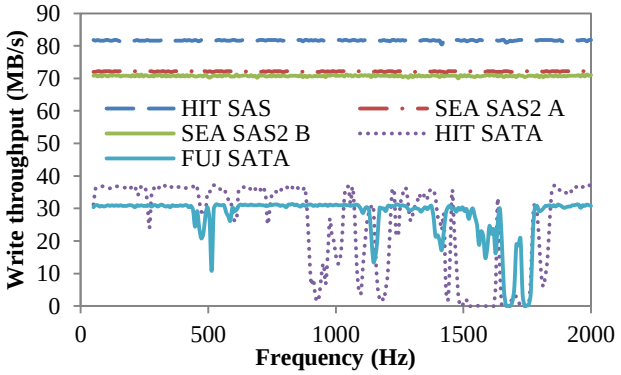

Figure 5. Throughput dependence on vibrational frequency, with amplitude fixed at 0.17g

4. MANAGEMENT POLICIES

Here we describe state-of-the-art Dynamic Load Balancing (DLB), two proposed thermal management schemes – "Joint Energy, Thermal and Cooling Control" [10] and "Application-level Oracle" [2] – in addition to our novel policy, "Full System Thermal Management". Whereas disk performance optimization is generally neglected by the conventional schedulers, or at best, included as an afterthought, our policy considers disk performance as one of the key factors in scheduling decisions. Figure 6 summarizes the data and control flow among system components for policies we evaluate.

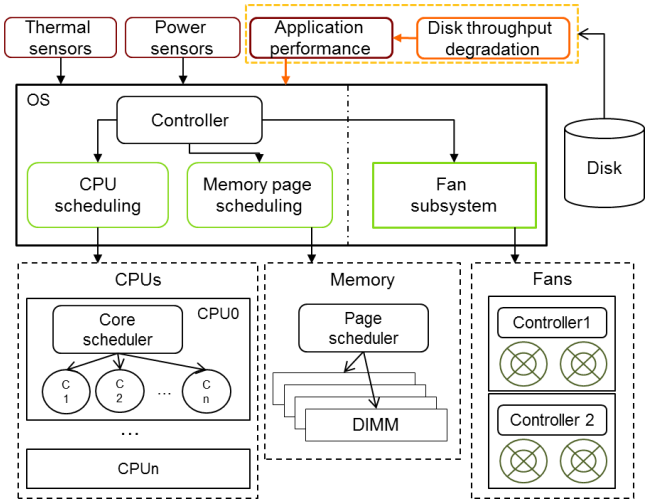

Figure 6 Overview of system control framework

A centralized controller is implemented as a finite state machine triggered by thermal and power sensors and tachometer feedback from the fans. The controller outputs the target power distributions for the CPUs and memory and a target temperature for the cooling system. The independent actuators in CPUs, memory and fans respond through workload scheduling, page migration and fan speed control respectively. The exception is for DLB, which has an independent fan controller responding directly to temperature sensors. In all policies, clock gating is used as a protection mechanism against very high temperatures. If a core (or memory module) temperature reaches the emergency threshold (e.g. 90°C in our case), hardware will increase the amount of time the core spends gated (e.g. in this paper 2% at each scheduling tick). Once temperatures return to acceptable levels, core gating time is decreased by 2% at each tick.

Dynamic load balancing with independent fan control (DLB): This policy implements power management and cooling management independently. In software, it balances task queue lengths among cores through task migration to manage power consumption and maximize system resource utilization. A separate proportional-integral (PI) controller sets the fans according to on-board temperatures. PI control operates based on a feedback loop reporting the present error (temperature reduction needed) and accumulated past errors (an integral of past temperature reductions), and its response time is tuned at design time. There is no power- or thermal-aware management of memory pages. DLB represents the state of the art cooling policy deployed in today's servers.

Joint energy, thermal and cooling control (JETC)[10]: This scheme centralizes the thermal management of CPUs and memory and fan subsystem, aiming to reduce aggregate energy consumption of the fans and memory modules. The controller assumes sensors for per-core power and temperatures. It proactively migrates tasks between cores, migrates pages between memory modules, deactivates idle memory modules and equalizes fan speeds among different controllers. This policy does not consider the performance or energy impact of disk access.

Application-level oracle (App-Oracle) [2]: This policy adds a contingency plan to JETC during periods of high I/O utilization. It is called an "oracle" as it relies on the application to provide exact information about I/O demand ahead of time, which is normally not available to applications or the scheduler. Constant monitoring of disk throughput is also required, but any system measurement overheads are neglected. Batch jobs are registered as low priority with the task scheduler and service jobs are registered as high priority. It performs integrated workload scheduling, and fan control. When an application forecasts a high demand for I/O (*iowait* time above 50%), the scheduler directs lower priority batch jobs to be cooled through DVFS. At each scheduling tick, it progressively steps down the clock speed of cores running batch jobs until the target temperature is reached, at which point it is progressively steps ups the clock again at each tick. This effectively trades batch job performance for lower temperatures and lower fan power consumption.

Full System Thermal Management (FSTM): We improve on the most recent vibration-aware policy [2] by making vibration-awareness part of the core philosophy. It does not rely on I/O characteristics to be self-reported at the application level, or intrusive system calls to monitor I/O current and near-term utilization. It both compensates for internally generated vibrations and provides resilience against external vibrations.

The fan-disk characterization curve (such as Figure 2) is segmented into either temperature-driven zones or disk-performance-driven zones according to the slope (m) of the curve. At moderate temperatures and low fan speeds, the slope is flat (m=0), indicating that I/O throughput is independent of fan speeds. Thus, fan speeds should be dictated only by the highest measured on-die temperature. At higher fan speeds, the slope becomes more negative. As long as the slope is above the threshold th= -1, it is considered temperature-driven by default. If the slope decreases past the threshold (m<=-1), the policy enters a disk-driven zone, identifying times where disk sensitivity is particularly high relative to the fan speed. Then, the guideline for

balancing fan speeds is relaxed in order to reduce vibrations. Fan speeds are set as a digital signal discretized from an analog measurement of temperatures, and air flow efficiency decreases with higher fan speeds. At thresholds between some fan step *n* and *n*+1, the controller assigns the minimum fan speed to obtain a higher gain in performance, with lower cooling capabilities.

FSTM also dynamically allocates a subset of SSD storage to serve as a buffer cache, in order to compensate for the expected performance overhead of fan vibrations. We execute 22 TPC-H queries [14] with recommended SSD cache allocations without thermal management optimizations, to estimate the benefit of turning on this caching feature. We make no assumptions about specific application behavior, relying on the database software features (e.g. Oracle 11g Smart Flash Cache [17]) to identify which frequently-used tables or indices should be stored in the SSD. In some cases, SSD caching shifts the performance bottleneck (and energy inefficiency) from I/O throughput to the CPU, so using SSDs is not a guaranteed performance booster without detailed optimization at different system layers, including device drivers, firmware and database libraries. In general, it increases speeds by up to 40%, but in some instances it adds up to 9% performance overhead. The policy selects between 1%, 2% and 10% of the database table size to allocate for SSD caching, shifting reliance away from spinning media towards the SSD as fan speeds increase. When the fan speed is below 50% PWM, the cache size is 1% of the table size; when it is between 50%-75% PWM, the cache size is 2% and when fan speeds are above 75% PWM, up to 10% of table size is allocated.

Lastly, much like with the other state of the art policies, the centralized controller uses task migration, page migration, and fan control. These decisions are made proactively based on a band-limited temperature predictor (BLP) [13].

5. RESULTS

We evaluate the management policies with a mixed workload of data or memory-intensive service jobs and compute-intensive batch jobs. We use commodity SATA disks as they are preferred by cost-sensitive datacenters for their low cost per storage density, augmented with a single SSD drive per system. Buffer caches are enabled to capture the real response of applications along with power, thermal, cooling and disk performance issues.

TPC-H is a decision support benchmark representing databases requests [14]. The queries comprise combinations of operations such as sequential scan, index scan, merge join, and hashing functions. We choose *query 1, 3, 10, 13 and 19* to represent data-intensive service jobs. SPEC CPU2006, on the other hand, is a benchmark suite targeted towards compute-intensive workloads [15]. Of the SPEC suite, we chose *bzip2*, a compression algorithm, and *hmmer*, a compiler, to represent batch jobs. In each workload set (Table 2), the processor is at 75% utilization, running a single TPC-H query and five SPEC tasks. With this mixed workload, we expect to encounter both thermal issues due to heavy computation, and I/O performance issues due to reliance on the disk access rates.

We used Smart Flash Cache available in Oracle Database 11g to measure the performance effect of allocating different amounts of flash cache. Since dynamic cache resizing is not exposed, we interleave performance traces taken from runs with different cache sizes specified at database initialization time. We monitor sensor statistics and event logs through IPMI. Disk access statistics were collected through *iostat* reports, estimating the number and average service times of queued and active transactions per sampling interval (every second). We use a modified version of HotSpot [16] to model the thermal interactions between the processor (with CPU cores, L3 caches and a crossbar), memory modules and the fan subsystem. We model a server with 8 cores running at 2.85GHz, with 8 DIMM modules of 16GB each. The database is configured with 1GB RAM per instance, with a variable flash cache size between 500MB – 10GB. OS scheduling is done every 1ms. Since the packaging thermal time constant is on the order of seconds, the temperature prediction distance is set to 9ms to allow accurate thermal-directed task migration decisions. The fan control interval is set to 1s.

Table 2 Job combinations evaluated

Workload ID	Benchmarks
W1	*5 bzip2, 1 tpchq1*
W2	*5 bzip2, 1 tpchq3*
W3	*5 bzip2, 1 tpchq10*
W4	*5 bzip2, 1 tpchq13*
W5	*5 bzip2, 1 tpchq19*
W6	*3 bzip2, 2 hmmer, 1 tpchq1*
W7	*3 bzip2, 2 hmmer, 1 tpchq3*
W8	*3 bzip2, 2 hmmer, 1 tpchq10*
W9	*3 bzip2, 2 hmmer, 1 tpchq13*
W10	*3 bzip2, 2 hmmer, 1 tpchq19*

5.1 Performance improvements

For each workload, we inspect the execution time normalized to the application performance with Dynamic Load Balancing (DLB), shown in Figure 7. JETC [10] improves over the default since it lowers fan speeds, which naturally mitigates the internal vibrations. Only App-Oracle reactively trades off batch job performance for lower temperatures while minimizing times for the service jobs. This gives close to perfect disk performance relative to fan vibrations, but relies on oracle knowledge of I/O demand, which is an unreasonable requirement of schedulers. The dotted line shows how batch jobs may be penalized down to 90% of original speed in the case of workload 6. FSTM depends on a fine-grain profile of disk throughput and fan speeds, and a coarse-grain profile of flash cache sizes and overall application performance. FSTM-NoSSD achieves on average 51% speedup by optimizing disk throughput only based on the disk sensitivity to fan speeds, without SSD caching enabled,. With SSD caching, FSTM achieves a 59% speed up over DLB on average. It is up to 73% faster than DLB without incurring any penalty to batch jobs, unlike the App-Oracle which only delivers up to 59% speedup. Query 13 has a single phase of hash join, dominated by large sequential accesses [18], the ideal target case for SSD cache prefetching optimization. It clearly shows the most benefits from FSTM in W4 and W9 as more SSD cache is allocated at high fan speeds. In some cases, e.g. W5 and W7, a naïve use of SSD-caching, without additional database-level configurations, actually leads to more performance overhead than benefits, so FSTM-NoSSD performs better than FSTM. Though we did not inject the effect of external vibrations, FSTM already shows improvements over current policies. In a datacenter environment with external vibrations, it should show even better results compared to other policies, as the extra layer of SSDs reduces the system's reliance on spinning media.

5.2 Energy savings

Figure 8 shows energy savings within the socket (including cores, L3 cache and the crossbar), memory modules, and fans. JETC improves memory power consumption by limiting the subset of activated DIMM modules at the expense of slightly higher fan

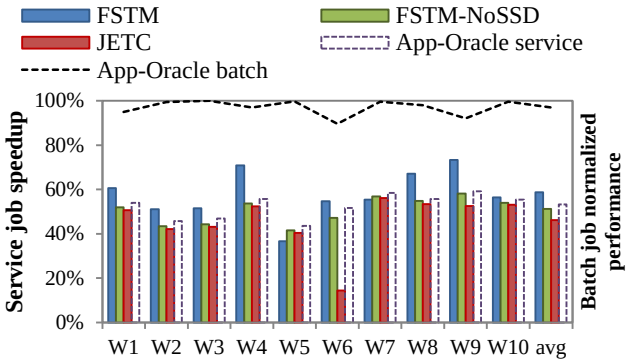

Figure 7 Service job speedup and batch job penalties (App-Oracle only) relative to Dynamic Load Balancing

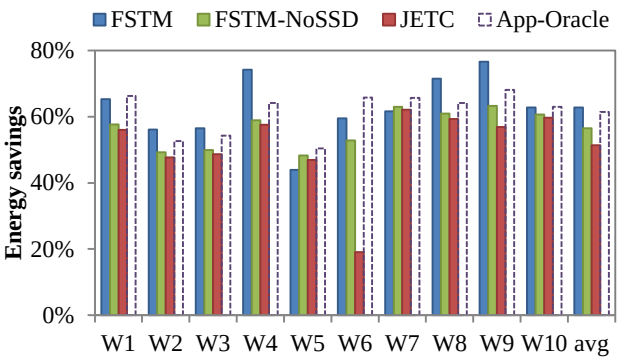

Figure 8 Energy savings in the socket, memory and fans relative to Dynamic Load Balancing

speeds. It neglects the disk performance, resulting extended system uptime and limited energy benefits. The three primary targets of energy consumption reduction were leakage power due to high temperatures, the fan subsystem, and system uptime power draw. The extended system uptime may come from core and page migration delays, DVFS of CPU-intensive jobs or emergency core gating, but the largest component comes from disk access delays. For these measured workloads, while CPUs do spend more time idling for disk access to return during high fan speeds, we observe that dynamic power consumption of a core decreases only if the disk throughput drops below 40% of its ideal. FSTM reduces peak and average core temperatures, lowers fan speeds and minimizes job execution times through fan optimization and the use of flash caching; thus we achieve an average of 63% and maximum of 77% in energy savings.

6. CONCLUSION

As on-chip power densities increase and workloads comprise more response-time-critical service jobs, it is crucial to develop efficient, performance-aware thermal and cooling management techniques. In this work, we present a systematic characterization methodology for enterprise servers, quantifying the unexpected effects environmental vibrations have on disks. Our characterization aids in formulation and evaluation of scheduling policies, and may also inform datacenter design. We propose a realistic cooling and thermal management policy where consideration for disk performance is a key factor in scheduling decisions. Our policy integrates CPU and memory modules scheduling, cooling, and mitigation of fan-induced performance degradation in a way that is implementable in real systems. We also capture the trend of hybrid SSD-HDD storage systems to improve runtime performance and system resilience to vibrations. This solution combats both internal and external vibrations via a simple firmware update and one additional SSD per system. By allocating SSD temporary storage for only up to 10% of the database table size, we achieve up to 73% speedup in job performance and energy savings of 76% over current schedulers and controllers.

7. ACKNOWLEDGEMENTS

This work is supported in part by Oracle Labs, National Science Foundation grant nos. 916127, 1029783, and 821155, DARPA, MuSyc Center, SRC Global Research Collaboration Task 2169.001, and the UC San Diego Center for Networked Systems.

8. REFERENCES

[1] Lefurgy, Charles, et al. "Energy management for commercial servers."*Computer* 36.12 (2003): 39-48..

[2] Chan, Christine S., et al. "Fan-speed-aware scheduling of data intensive jobs."*Proceedings of the 2012 ACM/IEEE international symposium on Low power electronics and design*. ACM, 2012.

[3] Szabados, David. "Are All Hard Drives Created Equal? Examining Rotational Vibration in Desktop vs. Enterprise." <http://enterprise.media.seagate.com> 2010.

[4] Narayanan, Dushyanth, et al. "Migrating server storage to SSDs: analysis of tradeoffs." *Proceedings of the 4th ACM European conference on Computer systems*. ACM, 2009.

[5] ASHRAE, TC. "9.9 (2011) Thermal guidelines for data processing environments–expanded datacenter classes and usage guidance." *Whitepaper prepared by ASHRAE technical committee (TC)* 9 (2011).

[6] Wang, Zhikui, et al. "Optimal fan speed control for thermal management of servers." *Proc. IPAC* (2009): 1-10.

[7] Donald, James, and Margaret Martonosi. "Techniques for multicore thermal management: Classification and new exploration." *ACM SIGARCH Computer Architecture News* 34.2 (2006): 78-88.

[8] Choi, Jeonghwan, et al. "Thermal-aware task scheduling at the system software level." *Proceedings of the 2007 international symposium on Low power electronics and design*. ACM, 2007.

[9] Skadron, Kevin, et al. "Control-theoretic techniques and thermal-RC modeling for accurate and localized dynamic thermal management." *High-Performance Computer Architecture, 2002. Proceedings. Eighth International Symposium on*. IEEE, 2002.

[10] Ayoub, Raid, et al. "JETC: Joint energy thermal and cooling management for memory and CPU subsystems in servers." *High Performance Computer Architecture (HPCA), 2012 IEEE 18th International Symposium on*. IEEE, 2012.

[11] Meza, Justin, et al. "Tracking the power in an enterprise decision support system." *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*. ACM, 2009..

[12] Patterson, Michael K. "The effect of datacenter temperature on energy efficiency." *Thermal and Thermomechanical Phenomena in Electronic Systems, 2008. ITHERM 2008. 11th Intersociety Conference on*. IEEE, 2008.

[13] Ayoub, Raid Zuhair, and Tajana Simunic Rosing. "Predict and act: dynamic thermal management for multi-core processors." *Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*. ACM, 2009.

[14] TPC Benchmark H <http://www.tpc.org/tpch/> 2012.

[15] SPEC CPU2006 <http://www.spec.org/cpu2006/> 2012.

[16] Huang, Wei, et al. "An improved block-based thermal model in HotSpot 4.0 with granularity considerations." *University of Virginia* (2007).

[17] Oracle Exadata. "A technical overview of the Sun Oracle Exadata storage server and database machine." <www.oracle.com/us/solutions/datawarehousing/039572.pdf> 2009.

[18] Kandaswamy, Meenakshi A., and Robert L. Knighten. "I/O phase characterization of TPC-H query operations." *Computer Performance and Dependability Symposium, 2000. IPDS 2000. Proceedings. IEEE International*. IEEE, 2000.