GreenLight Technical Report

2009

Power and Thermal Management

Tajana Simunic Rosing: Power/Thermal Management

Research Activities

The goals of our work are to develop strategies to analyze the power, thermal and workload dynamics, and to design management methods to reduce power consumption while mitigating the temperature induced problems in datacenter environments. Thus far we have implemented and tested a number of reactive and proactive thermal management techniques evaluated the benefits of optimizing temperature while minimizing energy versus only focusing of lowering the system power/energy consumption. We are in process of extending Xen, a virtual machine monitor for IA-32 IA-64 and PowerPC 970 architectures, to enable power and thermal management within the virtualization technology.

Comparison of power and thermal management
A number of dynamic power management (DPM), dynamic voltage scaling (DVS) and thermal management scheduling policies have been proposed in the past but none consider trade-offs among power, reliability and the workload distribution on chip. In our recently published work [CoskunTVLSI2008] shows how to determine an optimal schedule for a given set of tasks/threads so that the deadline, dependence and reliability constraints are met, while the energy consumption is minimized. After formulating and using the integer linear programming (ILP) solver to get a solution to the scheduling problems for an 8-core UltraSPARC T1, we found through implementation that processor hot spots, thermal and spatial gradients can be significantly reduced (as much as 60%) while meeting the performance constraints and minimizing the overall energy consumption. Furthermore, only minimizing the energy consumption leads to a suboptimal thermal response. This is largely due to the fact that minimum energy is obtained by clustering the workload into as few as possible processing cores, thus resulting in localized hot spots, while the rest of the cores are placed to sleep, and therefore are significantly cooler. Adding thermal constraints completely solves this problem. Although the ILP solution is optimal, it is also static in that we need to know the workload ahead of time. Since dynamic thread scheduling for multicore CPUs is an NP complete problem, we have designed heuristic techniques to handle both power and thermal management jointly.

Online learning for power and thermal management
A number of heuristic and stochastic policies have been proposed in the past. Simpler DPM policies like timeout and predictive policies decide to go to sleep heuristically with no performance guarantees, while more sophisticated stochastic policies guarantee optimality for stationary workloads. It is very difficult to devise a single policy solution which guarantees optimality under many different workload conditions. Hence, we propose a novel setup for power and thermal management, where we maintain a set of policies (suited for different workloads) and design a control algorithm that selects the best suited one for the current workload. We formulate power and thermal management problems as ones of workload characterization and selection, and solve them using the online learning algorithm. The selection is done among a set of experts, which refers to a set of DPM policies and voltage-frequency settings, leveraging the fact that different experts outperform each other under different workloads and device leakage characteristics. The online learning algorithm adapts to changes in the characteristics, and guarantees fast convergence to the best performing expert.

We performed experiments on a hard-disk drive (HDD), Intel' PXA27x CPU, AMD's Opteron and Intel's Xeon CPUs with real life workloads. Our results show that our algorithm adapts really well and achieves an overall performance comparable to the best performing expert at any point

in time, with energy savings as high as 58% and 92% for HDD and CPU respectively [DhimanTCAD2009]. The evaluation of the CPU's temperature profile takes into account the hot spots, thermal cycles and spatial gradients. For this set of experiments we used an UltraSPARC T1 and datacenter workloads as measured by the Continuous System Telemetry Harness (CSTH) [CoskunTVLSI2008]. We achieve 20% and 60% decrease on average in the frequency of hot spots and thermal cycles, respectively, in comparison to the best performing policy, and reduce the spatial gradients to below 5%.

Proactive thermal management

Most of the previously proposed thermal management techniques focus on maintaining the temperature below critical levels and respond only once a given threshold is reached. Obviously, such reactive techniques maintain the temperature below a critical level at a performance cost. As a next step of our work we designed and implemented proactive thermal management methods that prevent thermal and reliability problems before they occur with negligible performance overhead.

In our experiments, we have seen that as the workload goes through stationary phases, the future temperature can be accurately estimated by regression of the previous measurements. Thus, we used autoregressive moving average (ARMA) for estimating the temperature given the recent history of data from the thermal sensors. Based on the temperature observed through thermal sensors, we predict temperature a number of steps into the future using with ARMA model. The scheduler then allocates the threads to units to balance the temperature distribution. During execution, the workload dynamics might change and the ARMA model may no longer be able to predict accurately. To provide runtime adaptation, we monitor the workload through the temperature measurements, validate the ARMA model and update the model if needed. To detect when updating is needed we propose to use the sequential probability ratio test (SPRT), which provides the earliest possible detection of variations in time series signals. In our experiments on an UltraSPARC T1 implementation, we have observed that our technique achieves 60% reduction in hot spot occurrences, 80% reduction in spatial gradients and 75% reduction in thermal cycles in average in comparison to reactive thermal management techniques [CoskunICCAD2008].

System monitoring and measurement strategies

Integrating energy and thermal management into the OS requires fusing together data collected from temperature sensors in the system available through IPMI interface, in addition to performance metrics related to the workload and processors. For the OS to make correct dynamic management decisions, it must take these disparate telemetry streams and reconstruct a logically consistent snapshot of the system state. Misaligned telemetry streams, in which correlated events occur out of phase, may cause instabilities in the management and scheduling decisions, and lead to suboptimal results. Prior work in temperature-aware scheduling has focused exclusively on simulation results, with little attention paid to how the proposed solutions would be implemented in practice. Continuous System Telemetry Harness (CSTH), developed by Sun Microsystems, systematically addresses the issues related to the collection, preprocessing, and analysis of time-series telemetry data [K. Gross, K. Whisnant, and A. Urmanov. "Electronic prognostics through continuous system telemetry" in *MFPT* 2006.]. It captures telemetry data from a variety of physical and logical sensors in the system. Physical sensors include distributed temperatures, voltages, currents, humidity, and vibration; and logical sensors include OS metrics, hardware performance counters, and quality-of-service metrics. CSTH is integrated into the existing software stack, and introduces negligible additional overhead. The CSTH mitigates the challenges of large scale remote monitoring schemes by providing a real-time telemetry architecture with a circular-file repository that acts as a system "black box" performance monitor. All monitored variables can be used for not only dynamic management, but also for analyzing and diagnosing

system failures. Telemetry data can be easily passed to the OS at regular intervals for energy- or temperature-aware job scheduling. CSTH enables enhancements to availability, serviceability, performance, capacity planning, quality of service, and security; but without placing burden on the monitoring infrastructure during the majority of the time when systems are behaving without problems.

Over this last period we ported and evaluated CSTH on a few Sun servers. The measurement results show that this tool is invaluable for providing data on temperature and workload characteristics at the whole server board level. We are currently comparing using CSTH versus tapping directly into on-board and on-chip sensors via the Intelligent Platform Management Interface **(**IPMI) and MSRs. Model specific registers (MSRs) of processors are a great way to dynamically read the core temperatures and performance counters. The readings are based on digital thermal sensors embedded inside the cores of the processor. In addition to temperature and performance, we have successfully instrumented and measured power at board level for all the important components in the GreenLight machine we have. The table below gives the measured power of the whole system (in watts) and the percentage contribution of the different components for a typical high end server (quad core Xeon in this case):

Components	**System Idle**	**System Active**
CPUs	28%	52%
Motherboard	14%	15%
Fans	66%	24%
HDD	3%	1%
Power Supply	22%	8%
Total (W)	**158**	**271**

The measurement gives insights into what can be done to improve the energy efficiency. The system has constant speed fans, which contribute 66% of power consumption during idle times. Effectively managing this can bring significant savings. The power supply is more efficient during active times (92% efficiency) than during idle times (78% efficiency). The power contribution of CPUs goes from 28% to 52% when the system is idle vs. highly utilized. Dynamic voltage scaling and power management algorithms can help reduce further the CPU power as outlined above.

Energy management in virtualized environments

Our current focus is on developing a novel power and thermal management framework called vGreen aimed at virtualized environments. The framework has been already implemented on GreenLight machines running Xen, and is able to provide insights into the power and performance requirements of workload dynamically at run-time.

We are currently experimenting with policies based on the information provided by this framework for energy efficient scheduling both within and across servers.

For further information please contact Tajana Rosing, trosing@ucsd.edu