CS800 Huangxin Wang G00773152

**Seminar 1: Content-Adaptive Display Power Saving in Internet Mobile Streaming**

* Speaker: Mengbai Xiao
* Date: 3/6/2015

In this talk, the speaker presents his work on designing and implementation of content adaptive display power saving mechanisms for reducing power consumption on mobile device. The main idea is to tune the backlight based on GPU rather than CPU, which allows them to design more complex algorithms. They designed an algorithm based on dynamic programming.

As the battery capacity is constrained in mobile devices, it is important to design energy efficient mechanism to increase the power-on time of mobile device. It shows that the display consumes 68% power during the playback, and most of the energy is consumed by back light. In order to save power, backlight scaling techniques has been proposed. This technique reduces power consumption by dimming the display backlight. Meanwhile, the brightness perceived by the human eye is maintained by increasing the affected image's luminance.

The strongness of their mechanism is that they solved the following three challenges. First, to maintain image fidelity, the backlight level cannot be lower than a point determined by the brightness characteristic of an image. This constraint requires that the maximum pixel luminance of every frame to be computed, which can be both time consuming and energy intensive. Second, it is infeasible to adjust the backlight level for every frame because the display hardware takes time to perform the adjustment. Third, larger inter-frame backlight variation would cause flickering effects.

Their technique is based on the fact that by backlight scaling and combined RGB values with the tuned backlight, the output could be the same, while the power consumption reduced. In their implementation, YUV-RGB conversion is required. They use GPU rather than CPU as GPU is better at computation. However, GPU is also energy consuming. They implement the designed mechanism on android and experiment with more than 470 randomly selected YouTube video clips. They dim the light a decision model to predict the power savings resulting from the CAD. The results showed that the designed mechanism can reduce energy consumption which not affects users' satisfaction. In their experiments, 83.4% video can save power.

The weakness of their work is that they didn't do the real power consumption measurements in the experiments. Instead, they use a decision model that correlates the power consumption with backlight. However, it would be more accurate to do the real experiment. In addition, it is also interesting to measure the overhead of computation.

**Seminar 2: Greening Data Centers: Past, Present and Future**

* Speaker: Ricardo Bianchini
* Date: 03/27/2015

In this talk, the speaker introduces the energy management in green data center. In specific, he introduces the cooling strategy and workloads scheduling strategy. The workloads in data center includes the following three types: 1) high-performance computing and data analytics (batch), 2) online services (interactive) and 3) online services (interactive).

Conventional DC cooling uses cold/hot aisles, the temperature of the cold aisle at 20c or below; significant potential for hot air recirculation; the cooling infrastructure consumes a lot of energy. Three key advances in cooling including 1) higher cold aisle temperatures 2) hot aisle containment to prevent recirculation 3)"free" cooling: user outside air in certain climates. Larger DC operators now achieve yearly PUEs 1.0-1.20.

The strongness of their works is the reducing of tail latency. This is very critical in data centers, especially for user faced services. It determines the machines you need. If you can use fewer servers to achieve the same performance, then you can save energy. Currently, the standard is to guarantee quality of 99% percentile latency. The point is that if you can improvement on latency of 20% percent, then you can reduce server numbers by half. They reduce tail latency by parallelism. Often, people can use intra-request parallelism: reduce latency, but may overload the system. In their approach, they design few-to-many parallelism and parallel the long request more severely, add parallelism at fixed interval, but no fixed period is available. In particular, short period is good at low load, long period is good at high load.

The weakness of their works is that they didn’t compare with existing work. In particular, there is one paper in SIGCOMM 2014 which also introduces using parallelism to reduce tail latency, the paper name is “[Decentralized Task-Aware Scheduling for Data Center Networks](http://research.microsoft.com/apps/pubs/default.aspx?id=215429)”. It would be better for the speaker to talk about their differences with existing works.

**Seminar 3: Practical, Automatic Analysis of Mobile Software using Rational Analysis**

* Speaker: Hamid Bagheri
* Date: 4/22/2015

In this talk, the speaker introduces his work on mobile software analysis. The motivation of his work is to build dependable software. Formal verification techniques include exploring state space, which is heavy weight because it requires all cases. Also it requires human assistance and not suitable for software modeling. Alloy is a light weight former method, it can automatically analyze based on SAT solving. It uses simulation and check assertions.

In the compositional analysis of android inter-app vulnerabilities, the app markets are good malware delivery system, among which Android is the primary target. Android provides a very flexible communication models, the main components is the basic building blocks of android apps, android uses intent for event message. The android permission model allow the install time permission request. However, due to the inadequate permission model, it may bring weakness. For example, the vulnerability include privilege escalation, conclude permissions.

The strongness of their work is that they beat the naïve approach, we combines apps to analysis, which has scalability issues. They use four steps of extract information for each app, and extract the principal entities and permissions for manifest file. Then they identify the intents and intent filter that are latent application bytecode, and they determine the event-driven behavior of app, and they identify the vulnerable paths with each component. The experiment results show that they can check hundreds of components in minutes and thus have better scalability.

The weakness of their approach is that their analysis of the apps is limited. For example, there are other types of vulnerability such as intent hijacking. Thus it would be better for them to extent their analysis of the vulnerabilities of apps to wider areas and cover various vulnerabilities. Besides, their analysis focus on Android, I am curious whether their mechanism will also be applicable for other mobile operating system.

**Seminar 4: Automatic, Optimal, and Near-Optimal Resource Allocation in Cloud Computing**

* Arwa Aldhalaan
* Date: 04/23/2015

In this talk, the speaker introduces the works that she has done during her Ph.D. The topic she works on is resource allocation in cloud computing. The cloud computing is based on large distributed system, and resource management in cloud is very challenging. The objective of the author is to manage the resource in cloud in dynamic environments. For example, such tasks will decide how to process customer’s requirements by allocation machines and the number and types of machines to run services.

The challenge of this research is that the service request is dynamic, and the mechanism will need to make online decisions. In particular, an automatic, optimization techniques to meet QoS requirements is desired. The architecture is that in the cloud. Users make request to SaaS, then SaaS requires IaaS, then IaaS will design how to migration jobs among different virtual machines, and how to migration VMs in different cloud.

The strongness of their work is that they considering live virtual machine migration. The pages of address of virtual machine are copied when the VM is running. Then they dynamically migrate the dirty pages. The advantage of this approach is that it incurs very low downtime. They design a heuristic algorithm to solve this problem. Another strongness is that they take the SLA into account. Two virtual machines in the same physical machines will have better communication qualities. In this problem model, customers can decide the service of quality they want, and pay fees in according to the service quality. They design mechanism to decide how to allocate virtual machines to users in dynamic manner. They mechanism uses hill climb heuristic approach.

The weakness of their approach is that they didn’t bound the performance of their mechanism. The mechanisms they designed are often heuristic based, although the problem is NP hard. They can formulate it as non-linear programming. However, as far as I see, to compare with the optimal approach, they just enumerate all possible approach to find the optima value. This approach limits their simulation skills.