Paper Critique – Week 11

Name: PHAM CONG VINH Student ID: 2019711010

Efficient Virtual Memory for Big Memory Servers

This paper was published at the time that high performance computers and servers get significantly increasing in memory when the price had being decreased by state of the art of manufacture. The problem is that virtual memory which is a historical consequence but induced overheads as memory intensive workload work in the way that totally not leveraged any benefit of virtual memory at all.

By conducting experiments, the author showed that the virtual memory translation operation, which facilitated the translation-lookaside buffers (TLBs) speedup technique, is the main reason for the overhead and limiting the overall performance. They proposed a solution hardware based "direct segment" concept which mapping the virtual addresses within the range of normal memory page size directly to physical address. This concept based on 3 characteristics which are (1) retains a standard linear virtual address space; (2) is not overlaid on top of paging and (3) coexists with paging of other virtual addresses.

The strength of this paper is that they argue the obsolete of decades-old page-based virtual memory by matching it with the problem of emerging big-memory workloads. Because this is practical needs so it easily succeeded in convincing audience follow their solution.

The limitation of this paper is it hasn't considered yet whether their approach can negatively affect workloads which already be aware the problem of TLB overhead and had their own implementation, or not.

Coordinated and Efficient Huge Page Management with Ingens

This paper is similar with the above paper in a way that it try solving the problem about overhead caused by TLB misses. Recently, operating system (kernel and hypervisor) and memory intensive application has been aware of degradation of virtual memory, "base page" (memory page with size in 4KB). They had their strategy by introducing "huge page" supports and their own implementation to overcome the problem. But in practical, these solutions aren't complete and tends to cause more trouble.

The proposed solution of author is Ingens, which is memory manager for operating system and hypervisor. The design philosophy of Ingens is based on two principles: (1) considering memory contiguity is an explicit resource and should be allocated in a reasonable way; (2) maintaining information about spatial and temporal access patterns in order to facilitate the prediction the way of memory contiguity allocation that get most profits.

The strength of this paper is by conducting experiment, the author justify that Ingens will not harm the performance of application which has their own solution for the problem but will be beneficial to general operating systems, applications which hadn't been optimized.

The limitation of this paper is that author mostly focused on Linux, some comparison with other operating system such as FreeBSD, Solaris would be more interesting.