IOLanes: ADVANCING THE SCALABILITY AND PERFORMANCE OF I/O SUBSYSTEMS IN MULTICORE PLATFORMS

Publishable Summary

Contact: Prof. Angelos Bilas, bilas@ics.forth.gr, FORTH-ICS, Greece

05-July-2013

The IOLANES project consortium is composed of the following partners:

- Foundation for Research and Technology – Hellas (FORTH), Greece – Coordinator
- Barcelona Supercomputing Center (BSC), Spain
- University of Madrid (UPM), Spain
- INTEL Performance Labs, Ireland
- IBM Research Labs, Haifa, Israel
- Neurocom S.A., Greece

IOLanes is funded by the EC under the 7th Framework program and is part of the portfolio of the Embedded Systems Unit - G3 Directorate General Information Society.

Executive Summary

Modern society is driven by an insatiable need to store and process information. Data is at the heart of every modern economic and social activity. Therefore, ICT infrastructures that enable processing of data are an important catalyst for innovation and growth.

In this context, infrastructure and workload consolidation in data centers is seen as a main enabler for improving the efficiency of modern ICT. Data centers amortize the use of infrastructure across applications, and they are already significantly improving our ability to process large amounts of data. Although the exact type and amount of processing we will need to perform on data in the future is not yet known, it is becoming apparent that current data centers' size and cost are limited by technology and cannot keep up with data growth. Since building larger infrastructures is not viable, it now becomes imperative that we carefully examine the efficiency at which our ICT infrastructures operate.

IOLanes has built fundamental technology for improving server efficiency and reducing overheads in modern data center servers. Workload consolidation in data centers, although it allows sharing of resources, by dictating the use of virtualization, makes application performance unpredictable. To avoid both problems, providers today operate servers at low utilization.

IOLanes has addressed these fundamental problems by redesigning the I/O path from application to devices in the Linux kernel. The IOLanes approach demonstrably improves application behavior by orders of magnitude for response-time sensitive workloads and by large factors for most workloads, under interference. In addition, the IOLanes mechanisms can effectively achieve bare-metal performance for modern, state-of-the-art virtualization systems.

IOLanes technology has already been licensed and is used in data centers. Equally important, however, is the understanding that IOLanes has developed of the factors that contribute to server (in)efficiency. We expect that this knowledge will impact future server design and will allow future data centers to achieve higher compute and storage efficiency, and to keep up with data growth.

Introduction

With increasing amounts of data and energy costs, infrastructure and workload consolidation in data centers is seen as a main enabler for improving the efficiency of modern ICT. Data centers amortize the use of infrastructure across applications and until now have significantly contributed to reducing ICT costs. Simply put, capital and operational expenses are reduced with economies of scale. However, it is becoming apparent that infrastructure size and cost are limited by technology and cannot grow at the same rate as data. Figure 1 shows that data will almost double every two years between now and 2020. In addition, emerging applications require more extensive processing over accumulating data. As eloquently stated in a recent report[1]: "We are running out of power and space".

Since building larger infrastructures is not viable, it now becomes imperative that we carefully examine the efficiency at which our ICT infrastructures operate. A report by Emerson Network Power[2] mentions that for x86 servers, typical utilization rates are between 7% and 15%, whereas energy use in low-performance servers can reach 60% to 70% of their nameplate rating. Therefore, increasing server utilization and reducing power consumption are the two main directions that can further scale our ability to store and process data and drive innovation.

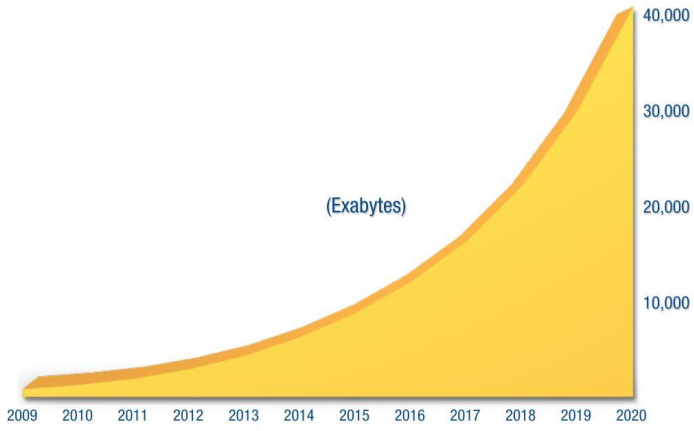

Figure 1: The digital universe: 50-fold growth from 2010 to 2020.

IOLanes has built fundamental technology for improving server efficiency by increasing utilization and reducing overheads in modern data center servers.

Workload consolidation in data centers, although it leads to more efficient use of the underlying infrastructure by sharing available resources across workloads, also leads to increased overheads and unpredictable application performance.

Workload consolidation dictates the use of virtualization. Virtual machines that host each application, and virtualization technology in general, create a well-defined interface between complex workloads and complex ICT infrastructures, and allow deployment and migration of applications anywhere in the data center. However, virtualization induces significant overheads, especially for data processing applications that require frequent access to large amounts of data. Virtualized workloads typically access data multiple times slower than applications executing natively on the same infrastructure.

In addition, consolidation of virtualized workloads on modern, multicore servers leads to workload interference. Applications trying to access the same pool of resources within servers, including memory, cores, storage, and network, interfere with each other, causing significant degradation in performance. As a result of workload consolidation, users may experience both reduced

[1] Technology Brief: Utilization and data center productivity. Virtualization and automation drive dynamic data centers. Computer Associates (CA), 2008.
[2] Data center Infrastructure Management. The Promises, Challenges, and Imperatives of Data center Infrastructure Management. Emerson Network Power, 2010.

performance and significant performance variations. For this reason, providers tend to keep server utilization low, in an effort to not impact user satisfaction.

Virtualization overheads and interference are the main current inhibitors for increasing server utilization and infrastructure efficiency for data-intensive applications.

IOLanes has addressed both of these fundamental problems by redesigning the I/O path from application to devices in the Linux kernel. IOLanes eliminates virtualization overheads and workload interference on modern multicore servers, effectively improving infrastructure efficiency.

- IOLanes has redesigned the I/O path in the hypervisor to allow for dedicating server resources to individual workloads and eliminate any interference across workloads at the memory and device levels. We show that our approach improves application behavior by orders of magnitude for response-time sensitive workloads and by large factors for most workloads.
- IOLanes has redesigned the host-guest I/O path in modern servers to eliminate overheads both in the issue and completion paths. We show that our mechanisms can effectively achieve bare-metal performance for modern, state-of-the-art virtualization systems.

Figure 2 shows that as server load increases, IOLanes improves application behavior by multiple times, allowing providers to operate servers at higher utilization without impacting application performance.

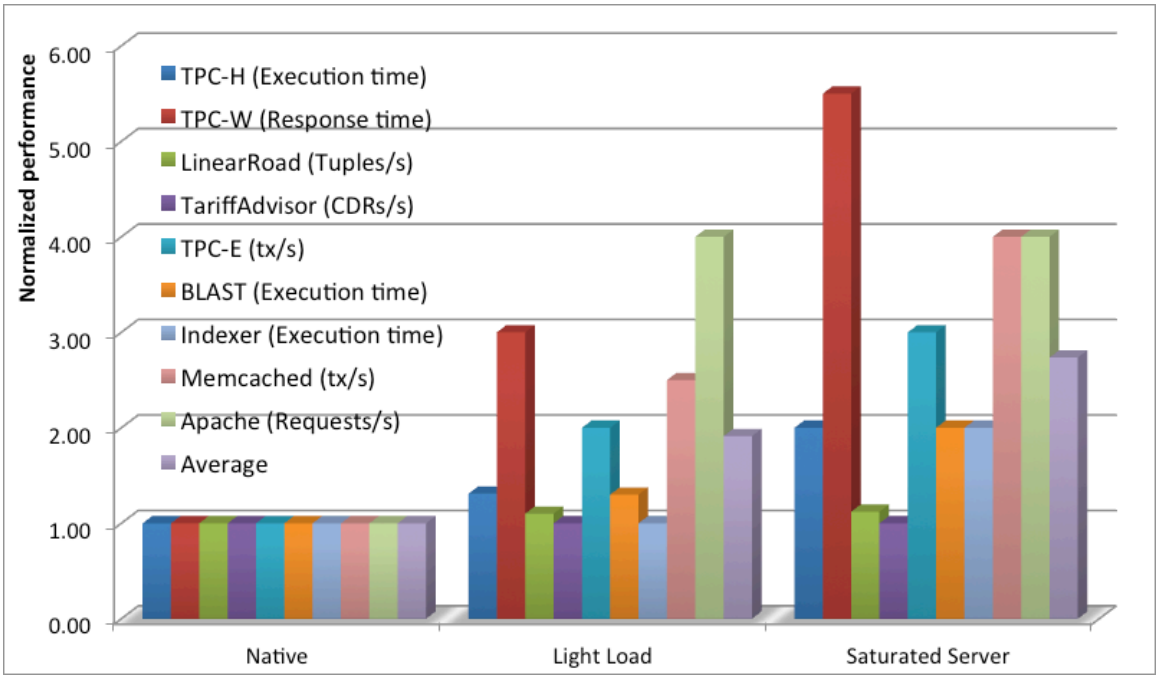

Figure 2: IOLanes improves application behaviour by multiple times as server load increases. The y-axis is improvement in application performance with IOLanes over native. As server load increases, IOLanes shows significant improvements over existing servers, allowing providers to operate at higher utilization levels without impacting application performance.

Next, we discuss the technology that has been developed by IOLanes and its impact on application performance and server efficiency.

Technology

IOLanes has examined fundamental issues related to I/O and multicore servers. Data center infrastructures are already starting to use large multicore systems for consolidating workloads to increase overall efficiency. This approach will only be effective if these systems achieve high utilization. However, providers tend to keep utilization low to avoid variations in performance and to cope with potentially high overheads during periods of intensive I/O.

This problem has not become critical until recently, because servers typically include a low amount of I/O resources, e.g., due to disk throughput. Today, however, and more so in the future, with GBytes/s and millions of IOPS available for I/O, sharing resources among 10s or 100s of workloads exacerbates overheads and contention. Many applications today already exhibit a sudden drop in performance when a backup application starts to read data from the same storage devices. Similarly, sharing the kernel I/O cache across workloads has been the source of performance unpredictability and degradation due to the global replacement policies. Virtualization introduces high overheads in the I/O path when crossing the expensive guest-host VM interface. Typically, applications have a significant drop in I/O performance when migrating from native to virtualized environments. In addition, device level scheduling is not able to cope with mixed workloads, given that each device only sees partial information about the entire workload. Both of these issues, inefficient sharing of resources and virtualization overheads, essentially lead to underutilization of resources in modern data centers.

Typical approaches in the past have tried to “manage” the sharing of resources mainly by appropriate scheduling. These approaches aim to maintain flexibility in sharing resources, in the sense that one workload can use all available resources in the system if there is no other demand. However, managing resources as a shared pool incurs overheads and interference in the management mechanism itself. In IOLanes we follow a different approach, where we partition resources for use by individual workloads. Each workload is allocated a “slice” of the I/O path and other workloads are not able to access these resources. Our approach reduces synchronization in the management mechanism, e.g., for shared buffers, enables placement of resources, such as I/O buffers on NUMA systems, and dramatically reduces interference of independent workloads, e.g., due to global replacement policies.

In addition, we have dramatically reduced the overhead of the guest-host interface for I/O, both in the issue and completion path, approaching bare-metal performance for many workloads. This allows virtualization to be employed without concerns for increased overheads during periods of high I/O load. Our main technique is to reduce the required VM exits when crossing the boundary between guest and host OS both in the issue and completion paths.

Finally, IOLanes has provided scheduling optimizations for mixed workloads in virtualized environments, by analyzing at runtime the I/O pattern and selecting the best-behaving scheduler dynamically.

We have prototyped this technology in a working system that is able to support real workloads and we have evaluated our approach with OLTP, OLAP, Streaming, and Web applications. We show a dramatic improvement, for instance reducing the “cycles per I/O” for TPC-E by two orders of magnitude compared to native execution, and by one order of magnitude compared to the existing Linux kernel mechanism for limiting use of resources. TPC-W response time improves by more than one order of magnitude with IOLanes technology when run in a consolidated server along with other applications. Virtualization overheads are practically eliminated allowing workloads to reach 97%-100% of bare metal performance. Our prototype includes a detailed monitoring infrastructure that has enabled us to observe in depth the behavior of the full I/O path in complex scenarios and with real workloads.

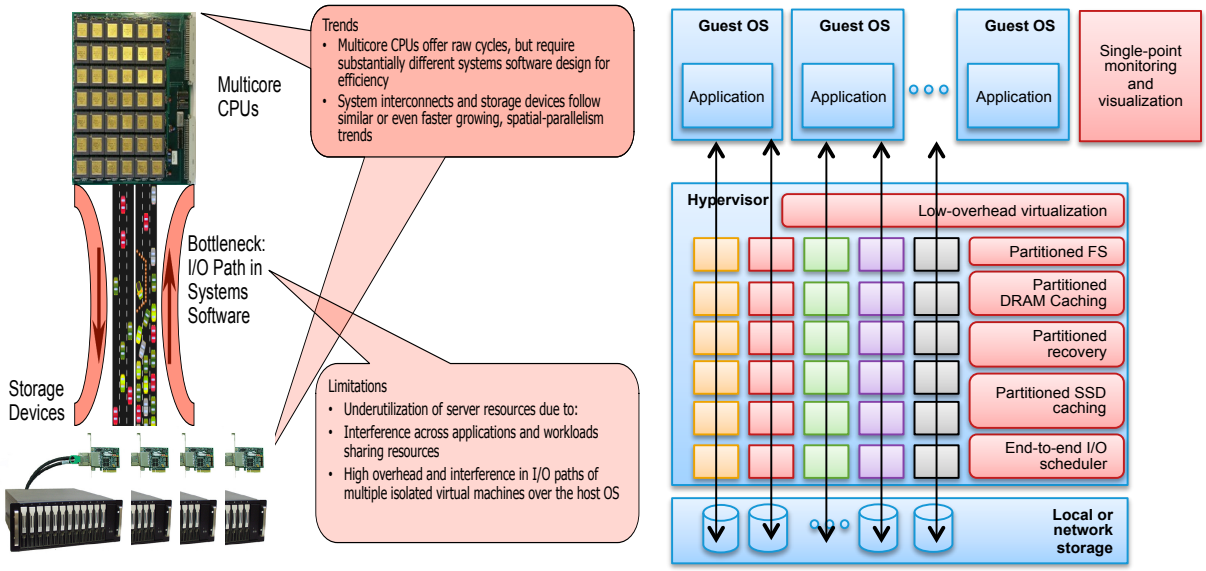

Figure 3: Traditional (left) and IOLanes (right) I/O path design. The I/O bottleneck in modern servers shifts from devices to the host stack, leading to interference, overheads, and eventually poor utilization of servers.

Overall, three high-level conclusions emerge from our results:

- **Analysis of server dimensioning for I/O:** With the current trends towards consolidation of resources, future servers will include larger amounts of CPU, memory, and I/O resources. Our analysis shows that servers for I/O workloads should be dimensioned to support 1 GByte/s/core of memory throughput, 1 GBit/s/core of I/O throughput, and 100K IOPs per core. Already at 512 cores, these are challenging targets to achieve. Especially for I/O, these numbers show that the use of NAND Flash (and other NVM technologies) is the only way forward to achieve the required I/O density.
- **The need for consolidation and virtualization:** What is more important is that servers with large amounts of resources will run multiple workloads to achieve high utilization and efficiency. Such workloads will have to run under virtualization to provide isolation across administrative domains. However, in this case overheads increase significantly when sharing server I/O resources across multiple virtualized workloads. For instance, when running multiple VMs in a single server, each I/O consumes 10x more CPU cycles already at 128 cores (projected). In addition, workloads interfere with each other while accessing their data via the shared I/O path, resulting in significant variability in performance.
- **Core I/O technology:** Our technology addresses two main problems in this landscape:
  (1) We isolate workloads from each other while accessing shared resources in consolidated servers, and achieve native performance in the presence of other workloads.
  (2) We reduce dramatically virtualization overheads associated to I/O in virtualized servers and achieve in many cases bare metal performance.
- **Real prototypes:** Our work and experience shows that building working prototypes and examining real applications introduces several complications, however, is necessary if we are to understand how future infrastructures should be designed to achieve the required utilization and efficiency.

For more information on the IOLanes technology please visit: http://www.iolanes.eu/innovation-technology.html

Potential Use and Impact

IOLanes has paved the way for improving the efficiency of servers in future data centers by contributing to three broad areas:

- **Research:** Our work in IOLanes has significantly advanced our understanding of I/O problems in data center servers and has already led to subsequent projects that will generalize our partitioning approach for all resources in the hypervisor for consolidated servers, aiming to significantly improve server utilization and thus efficiency by 2020. The work of IOLanes constitutes a reference point for efficient I/O in consolidated, virtualized servers.
- **Industrial relevance:** IOLanes technology has already started to be used commercially in data center. We foresee that as IOLanes technology matures, more of the techniques we have proposed will find their way in real products and will impact data center storage I/O and server design.
- **Collaboration:** IOLanes has formed strong partnerships between academia and industry to further pursue issues in data centers.

The final incarnation of our approach, as demonstrated by our working prototype, focuses on showing the negative impact of workload consolidation and how we can mitigate this in modern servers. Consolidation dictates the use of virtualization to isolate administrative domains, and requires efficient use of server resources to improve total cost of ownership. Our approach enables both: IOLanes technology allows large multicore servers to become dense, virtualized pools of resources and to serve efficiently more and more data-centric workloads, as demanded by the fast data growth we are witnessing. With IOLanes technology, providers will be able to operate infrastructures at higher utilization without impacting application performance and user satisfaction.

We have not only designed but also implemented a working prototype and evaluated our approach with complex real-life workloads, including transactional, OLAP, and streaming applications. Our prototype can be deployed at the hypervisor level, transparently to all applications using the system. Management tools take care of partitioning resources, for instance, when required by migration of a new application to or away from the server. Applications remain unmodified and use the same interface, while the management (or billing) system provides the required information for configuring resource partitions for each application.

Moving forward, the technology and prototype we have built in IOLanes will be an invaluable tool for expanding our understanding of I/O issues in modern servers and improving server design and the efficiency of data center infrastructures.

Conclusions

Data is projected to play a pivotal role in most emerging applications and services, supporting innovation in both economic and societal activities. However, modern ICT infrastructures are not able to keep-up with data growth. As a result, the total cost of ownership for ICT infrastructures does not scale with data size and is becoming prohibitive for small businesses that typically drive innovation. Given these scaling limitations, we now need to consider the efficiency at which infrastructures operate.

IOLanes has opened new venues for improving the efficiency of consolidated servers and data center infrastructures. We have developed fundamental technology that allows multiple applications to share resources within large, multicore servers, without suffering from the detrimental effects of interference across workloads. Our technology achieves bare metal performance under virtualization for data-intensive applications, and improves application behavior by up to orders of magnitude under heavy server load for response-sensitive applications, allowing both users and providers to benefit from consolidation.

IOLanes technology has already been licensed and is used in data centers. Equally important, however, is the understanding that IOLanes has developed of the factors that contribute to server (in)efficiency. We expect that this knowledge will impact future server design and will allow future data centers to achieve higher compute and storage efficiency and to keep up with data growth.