

# Characterizing Performance in Virtualized Execution

Hussam Mousa<sup>†\*</sup>   Kshitij Doshi<sup>‡</sup>   ElMoustapha Ould-Ahmed-Vall<sup>‡</sup>

<sup>†</sup>Computer Science Department  
University of California, Santa Barbara  
Santa Barbara, CA 93106

<sup>‡</sup>Intel® Corporation  
5000 W Chandler Blvd  
Chandler, AZ 85226

husmousa@cs.ucsb.edu, {Elmoustapha.ould-ahmed-vall,kshitij.a.doshi}@intel.com

Workload execution in virtualized machines is rapidly becoming commonplace while support for self-virtualization is already available in mass market processors. From a performance analysis and prediction standpoint, virtualization introduces new sources of uncertainty in reasoning about the factors that impact overall system performance and efficiency. This paper describes a general methodology for profiling the vertical software and hardware stack of a virtualized platform. A key contribution of this paper is an instrumentation approach that enforces alignment between measurements of potentially correlated factors, so that statistical analysis techniques such as regression modeling can be used effectively against the data. The paper shows, through the use of preliminary examples, how one can use such instrumentation to obtain decompositions of the CPI (cycles-per-instruction) metric for two workloads when each is executed first on native hardware and then in an identically configured virtual machine.

## 1 Introduction

In recent years a trend toward virtualization of machines has accelerated in enterprise data centers and other high volume processing environments [21]. This trend is driven in large part by business needs to transcend such rigidities of static deployments as power, space, and thermal footprints, while minimizing wasteful energy use by redistributing execution among fewer machines than are needed during peak demand periods [18, 21]. Vendors of popular processors have delivered architectural support for processor self-virtualization by removing the need for software based execution shadowing or emulation, further accelerating this trend [1].

With virtualization entering the mainstream, it has become important to characterize its effects on workload performance, so that machines can be configured aptly and optimized for virtualized execution. It is also necessary to obtain profiles of platform behaviors under virtualized execution so that future modifications of software and hardware can take due account of factors that limit the performance of software when it is deployed inside virtual machines.

In particular, because virtualization gives rise to additional pressures on micro-architectural resources and introduces new types of behavior that affect performance, it is essential to identify the hardware and software factors whose impacts dominate performance and to calibrate those

---

\* Work on this publication was done while employed as a Research Intern at Intel® Corporation

impacts. For instance, the ability to relate increases in virtual machine instances on a given physical platform to changes in the demands upon the resources of the physical platform helps identify efficient load balancing policies when an ensemble of virtual machines is scheduled across a collection of physical machines. Similarly, a detailed breakdown of the overheads that arise when a workload's execution is virtualized provides an opportunity to optimize the workload, perhaps along with hardware configuration, so that the overheads are removed or reduced.

For non-virtualized - or native - execution, well matured performance profiling instrumentations at multiple levels of the hardware and software hierarchy, supported by analysis and modeling techniques, has provided a rich source of understanding for workload execution. Simulations of instruction traces from native execution have been another source for obtaining critical modeling parameters such as per instruction overheads arising from dynamic conditions that generate execution stalls. By comparison, instrumentation and analysis techniques for virtualized execution are significantly more incipient. Ironically, because virtualization introduces many dynamic behaviors - some through new usages of hardware such as co-execution of heterogeneous operating systems - its performance characterization demands high resolution yet minimally intrusive instrumentation. Likewise, its simulation requires a much broader range of instruction traces than what would be considered sufficiently representative for capturing native execution behavior.

This paper describes an approach for bridging the profiling instrumentation gap between native and virtualized execution, and shows by example a preliminary method for isolating the key contributors of performance overhead in both native and virtualized machine instances. As a first step, we apply workloads with single state behaviors in this study, but the methodology advanced in this paper is designed to apply equally to behaviors that are not steady over time. An aligned collection of hardware monitored as well as hypervisor (also known as the Virtual Machine Monitor) instrumented event counts is achieved with hypervisor embedded code, where the width of a single interval of event collection can be based on time, instruction count, or some other unit such as numbers of branches retired or intrinsic application phases. The triggers are programmatically generated either from the processor, or from within the workload that is of interest. Instrumentation efficiency is obtained by keeping control within the hypervisor to the maximum extent possible. This frugality and flexibility makes it possible to get fine grained intervals of hardware and software event data where the statistical relationships among events are captured within the interval boundaries. Each interval contains event statistics gathered from the hardware, the Virtual Machine Monitor, the Kernel and possibly the application itself. The paper illustrates how regression techniques can be applied to this data to gain key insights in the effects of virtualization on workload performance. We also demonstrate how the data can be used to quantify how the occurrence of various micro-architectural and virtualization events affects virtualization overhead.

## 2 Background

Virtualization has been used in various forms for several decades. Recently the renewed interest in virtualization produced two dominant models: *Para-Virtualization* and *Hardware Virtualization* [21]. *Para-Virtualization* runs on unmodified hardware and relies on the Virtual Machine Monitor (VMM) to manage the memory, allocate resource, and isolate the various guest kernels. It works by replacing critical calls in guest kernels with "para" calls to VMM equivalents. The VMM is then responsible for guaranteeing that calls comply with system policies and that guests are protected from failures by other guests.

Hardware Virtualization delegates many of the functions of the VMM to the processor. Guest kernels no longer need to be modified and can operate as if they are running on dedicated hardware. The guest kernel runs at slightly limited CPL0, the same hardware ring that native kernels execute within, while a new, root privilege is created for the hypervisor. The processor is designed to trap into hypervisor whenever one of a limited subset of operations is attempted from non-root privilege level, in order to prevent guest kernels from piercing the abstraction of the processor that is presented to them by the hypervisor, or to affect the isolation of other co-executing kernels. Such traps, generally referred to as VT exits, are a common source of overhead in many of the early implementations of hardware virtualization, and therefore are the target of considerable optimization through successive generations of processor implementations and hypervisor algorithms. Virtualization of hardware requires virtualization of activities on the I/O path as well.

The primary difficulty that arises in fine-grained performance profiling in a virtualized context is in achieving completeness of performance state capture. A typical system can have multiple autonomous privilege domains in concurrent execution, where guest kernels do not exert direct control over each other (for profiling purposes) or over the processor. As the VMM is the only software component with control over the entire system hardware and memory space, it is the logical choice for basing a whole system profiler; yet, the VMM typically does not have sufficient visibility into guest encapsulated software performance state. Thus, an effective profiler must operate from the VMM where it has complete and direct access to the processor’s profiling features; and, its operation depends upon accurately allocating and aligning events to their appropriate domains. Guest kernel cooperation –even if to a very limited degree, also needs to be built-in so that guests can participate efficiently in triggering and pacing profile data collection and align activities at the workload level with performance data capture in the VMM at the desired granularity.

Finally, since the VMM is a major system level component, its actions will affect the overall workload behavior and the effects of that behavior on the processor. It is therefore important to include the VMM’s events as part of the whole system profile. In other words, to capture a comprehensive snapshot of the system’s behavior for any given execution interval, it is important to gather data from the processor, VMM, guest kernels, and executing workload. In the next section we present our proposed system design for a whole virtualized system profiler that addresses these challenges.

### 3 Virtualization Enabled Profiling

In this section we describe a profiling system we built to capture detailed behavioral snapshots of the entire hardware and software stack of a system –whether in native or virtualized execution. Subsection 3.1 presents the data model for our system, while subsection 3.2 describes the profiling infrastructure. We describe our experimental methodology and setup in section 4.

#### 3.1 Data Model

	CPI	PMC data			Hypervisor data			Guest Kernel data		
Interval <sub>m</sub>	CPI <sub>m</sub>	PMC <sub>1m</sub>	...	PMC <sub>nm</sub>	VT <sub>1m</sub>	...	VT <sub>nm</sub>	OS <sub>1m</sub>	...	OS <sub>nm</sub>

**Table 1: Data collection model for vertical virtualization profiling**

The basic unit of data in our system is an execution interval; a sequence of instruction delimited by an arbitrarily defined event. Intervals can be defined based on time slices, number of retired

instructions, or other micro-architectural or workload events. In this paper, we use intervals of equal number of retired processor instructions.

Each interval includes counters for events derived from three primary sources: hardware performance monitoring counters (PMC), guest kernels, and the hypervisor. The data model is depicted in Table 1. The system insures that events are synchronized and aligned to the intervals during which they were collected. Events are normalized to the number of retired instructions.

The collection of data at the granularity of an execution interval is motivated by the studies that demonstrate potential presence of varying performance behaviors during different parts of a workload's execution [22]. By collecting the data at the granularity of small intervals (millions of retired instructions), we are able to capture and account for these different behaviors. For a longer discussion on the rationale behind this method of data collection, the reader is referred to [19, 20].

### **3.2 Virtualized Profiling System Design**

To support the data model in Table 1, we developed a utility that achieves the following minimum objectives: aligned data collection from software and hardware sources, low overhead, and low application perturbation. The utility design is shown in Figure 1.

The system's primary controls are located within the hypervisor. They are responsible for configuring, starting, stopping, and reading all the various profiling agents. This controller receives a "recipe" from the user – a plan for an entire profiling session. This minimizes the need for any communication between the user's control system and the lower level profiling components during the profiling session. After configuring the various profiling agents, the controller synchronously starts profiling at all levels and waits for the trigger to end an interval. Upon receipt of the trigger, the controller again stops all the profiling agents in synchrony and reads the current event counts. These counts are stored in a buffer that is statically allocated within the hypervisor's address space to avoid data transfer during a workload's execution. It is important to note that the stopping and reading of the profiling agents usually occurs during a VT exit event so the guests are suspended anyway. We try to group these profiler's maintenance activities with other normally occurring VT exit events. This minimizes workload perturbation, since those VT exits are a "normal" part of virtualized system execution.

If the memory buffer for storing event frequencies fills up, a virtual interrupt is raised which is handled by the user level control utility. This causes a transfer of data from the hypervisor space to the user space, but it generally only occurs a few times –and only during the execution of really long workloads.

The PMC drivers are written to achieve maximum flexibility, efficiency and low overhead. Virtualization events (such as VT exits, fault injections, etc.) are collected through an interface with Xentrace [25]. Presently we capture counts of events, and we plan to capture more detailed information about specific events such as interrupt types and VT exit reasons in the future. Profiling agents that operate from guest kernels and capture guest performance state in synchrony will also be added to this framework in the near future.

Since guest kernels are in fact executing on virtual CPUs, a temporal mapping of virtual to physical CPUs is maintained for the purpose of proper designation of architectural performance events. Virtualization events collected are valid for the entire system, since they are effects of the hypervisor, a system wide component.

The native profiling version of our system is very similar to the above design except that the controller is located in the kernel instead of the hypervisor.

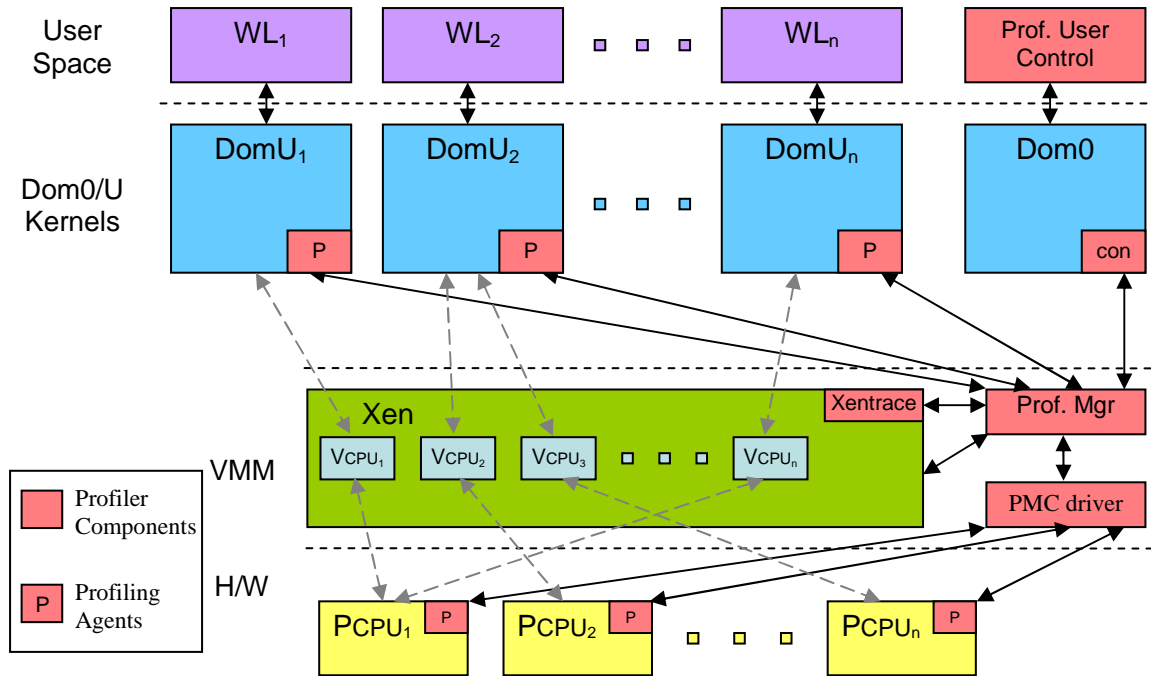


Figure 1: Virtual Profiler overall system design

## 4 Experimental Methodology

### Architectural Events

*L1 Instruction Cache Misses* [L1I\_MISSES]  
*L1 Data Cache Misses* [L1D\_MISSES]  
*L2 Cache Misses* [L2\_MISSES]  
*% Load Instructions* [LOADS]  
*% Store Instructions* [STORES]  
*% Branch Instructions* [BR\_RETIRED]  
*% Multiply Instructions* [MUL\_OPS]  
*% Divide Instructions* [DIV\_OPS]  
*DTLB Misses* [DTLB\_MISSES]  
*ITLB Misses* [ITLB\_MISSES]

*Page Walks* [PAGE\_WALKS]  
*Conditional Branch Mispredictions* [BR\_MISPRED]

### Virtualization Events

*VT Domain Transitions*  
*VT Interrupt Events*  
*VT Scheduling Events*  
*VT Page Fault*  
*VT I/O Event*

Table 2: Profiled Events

For this study, we collect 12 architectural events and 5 virtualization events. The events collected are listed in Table 2. Full descriptions of the architectural events are available in [12]. We run 10 benchmarks from the Dacapo java suite [4] release 2006-10-MR2 on the Sun® Java Hotspot™ Client Virtual Machine (build 1.4.2\_16-b05), and a custom benchmark consisting of building the Linux kernel image following random source file touches. Using the linux kernel compilation and randomizing the source files that are compiled creates a workload that exercises process creation and computation in some balance. Descriptions of the benchmarks are provided in Table 3. Each benchmark is executed on a native platform based on SUSE SLES10 SP1® 64-bit operating system and a virtualized platform based on the Xen hypervisor [3] and running an identical Linux guest kernel to the one in the native setup. Both the guest and the native kernel are configured with identical available memory, swap space and number of CPUs. The physical configurations are listed in Table 4.

<i>Dacapo</i>	
antlr	Parse a grammar and generate a parser and lexical analyzer
bloat	Optimize and analyze Java byte code
eclipse	Execute JDT performance tests for the Eclipse IDE
fop	Parse, format and generate a PDF file
hsqldb	Execute a number of queries against an in memory database
jython	Interpret a Python benchmark
luindex	Use lucene to index some books
lusearch	Use lucene to text search some keywords against a data set
pmd	Analyze a set of Java classes for a range of code problems
xalan	Transform XML documents into HTML
<i>Linux Build</i>	Touch several random source files and build a bzImage

**Table 3: Benchmark Descriptions**

	VT platform	Native Platform
Kernel	Linux 2.6.16.46-0.12	
Hypervisor	N/A	Xen 3.0.4_13138-0.40
Available Processors	2 Physical CPU	2 Virtual CPU
Physical Memory	1024MB (System limited)	
Swap file size	1024MB	
Hardware Platform	Intel® Core™ 2 Duo (dual-core) 2.66 GHz, 2 GB Main Memory, 4 MB L2 Cache, 32 KB L1 and 32 KB L2 cache	

**Table 4: Experimental Setup**

Due to the fact that there are only two reconfigurable performance counters on the Intel® Core™ 2 Duo, we use multiplexing as a technique for collecting more events than there are counters available. We group our events into sets that could be collected simultaneously. We then divide our desired interval width by the number of such sets, and use this new sub-interval width for the actual collection. We then switch the PMC collected events among the different sets at the end of every sub-interval. Finally we scale the collected PMC accounts back to the full desired interval width. Software events are not multiplexed but collected across the total number of sub-intervals that span an interval. For this study, we set the full interval width to 60 million retired instructions, and with six counter sets, the sub-intervals were set to 10 million retired instructions.

## 5 Results and Analysis

Based on the understanding of the system level operation of a virtualized platform, we can think of the performance as being impacted by two distinct – yet cross interfering – factors: additional architectural events due to virtualization, along with a potential increase in their costs, and additional virtualization (hypervisor) related events. Both factors will increase the Cycles per Instruction (CPI) metric, while the second factor will also increase the path length (the number of instructions per unit of work). In this paper we focus on the CPI and plan to study the impact on increased path length in future work.

In the remainder of this section, we describe several analytical approaches that can be used to gain insights into which factors contribute to virtualization overhead and the extent of their impact. Note that the actual values and relative weights presented in the next subsections are very specific to the processor micro-architecture as well as hypervisor and operating system versions, and will very likely decline rapidly as processor vendors deliver architectural improvements to reduce the cost and frequency of hypervisor transitions and interruptions. We believe, however, that our methodology and analytical basis are broadly applicable to the study of virtualization under any processor architecture, hypervisor, and guest kernel.

## 5.1 CPI Decomposition with Multi Linear Regression

Event Categories	Events	Linux/N	Linux/VT	Java/N	Java/VT
Instruction Cache	IL1_MISSES	1.36%	5.65%	4.41%	7.17%
Data Cache	DL1_MISSES	1.62%	5.15%	1.23%	2.88%
	L2_MISSES				
Load/Store Unit	LOADS	19.93%	24.72%	33.57%	25.29%
	STORES				
TLB group	ITLB_MISSES	16.48%	4.56%	7.71%	2.15%
	DTLB_MISSES				
	PAGE_WALKS				
Other	BR_MISPRED	60.62%	30.06%	54.07%	56.27%
	BR_RETIRE				
	DIV_OPS				
	MUL_OPS				
	CPI0 (constant coefficient)				
VT scheduling	VM Wake	0%	5.49%	0	1.39%
VT transitions	VM Enter	0%	20.69%	0%	5.04%
	VM Exit				
VT interrupts	VT Page Fault	0%	6.12	0%	0%
	VT Page Fault Inject				
	VT Exception Inject				
	VT Interrupt				

**Table 5: Grouping of Events and their relative contributions**

In this section, we describe a simple analytical approach to study virtualization overhead. It is based on the idea of CPI decomposition; that is to use regression analysis to statistically divide the CPI in its various components: e.g., core CPI, cache misses, branch misprediction, TLB misses and so on. This CPI decomposition is performed for both the native and virtualized execution. CPI decomposition also gives an estimate of how various VT events such as exits into the hypervisor affect the program execution time by contributing to an increased CPI. The

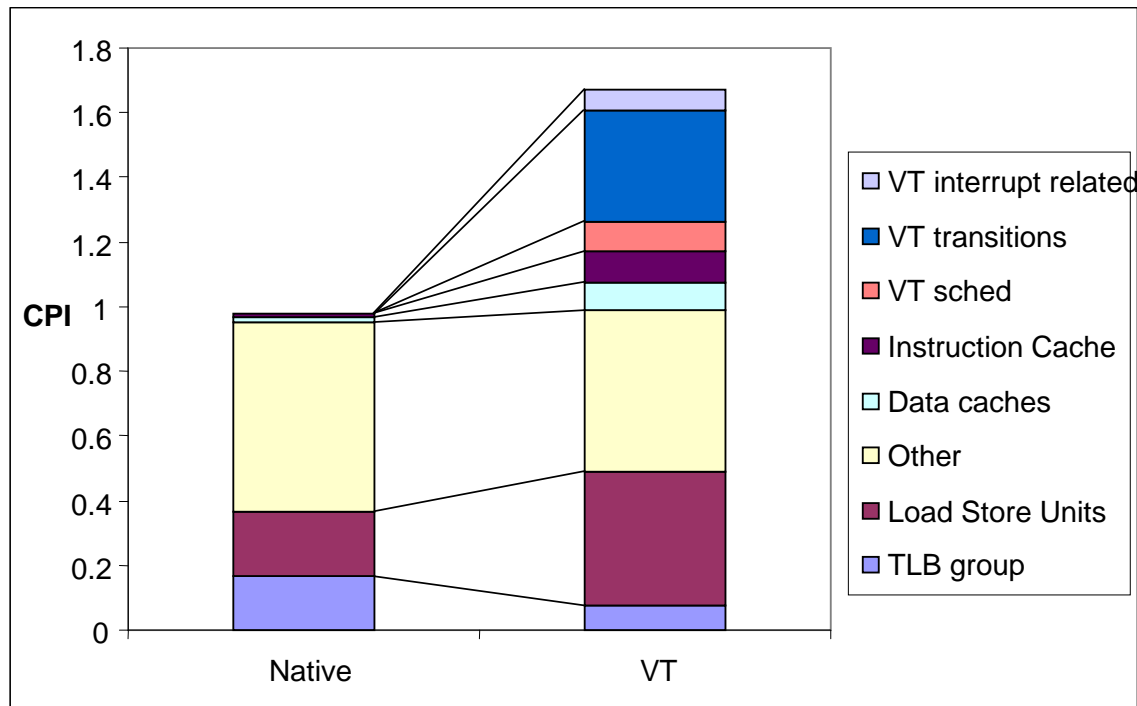
comparison between these two decompositions yields valuable insights into how virtualization changes the cost (in terms of CPI overhead) of micro-architectural resources.

We decompose the CPI by applying multi-linear regression with the CPI as the dependent variable, and the various architectural and virtualization events as the independent variables.

$$Y = \text{CPI} = \text{CPI}_0 + a_1 X_1 + a_2 X_2 + \dots + a_n X_n + b_1 V_1 + b_2 V_2 + \dots + b_m V_m$$

where  $\{ X_1, X_2, \dots, X_n \}$  and  $\{ V_1, V_2, \dots, V_m \}$  are the different micro-architectural and VT events, respectively. As expected, no virtualization event is used to fit the regression equation for native execution. The events used in this study are listed in Table 1. To reduce the effects of multicollinearity, we bundle the independent variables in groups of related events (e.g., TLB misses and page walks). The groupings and CPI contributions of individual events are shown in Table 5.

Figure 2 presents a visualization of the CPI decomposition for both the native and virtualized cases of the Linux build benchmark. By decomposing the CPI into the constituent components, we can observe the relative change due to virtualization and estimate how much of the overhead is due to the different events.

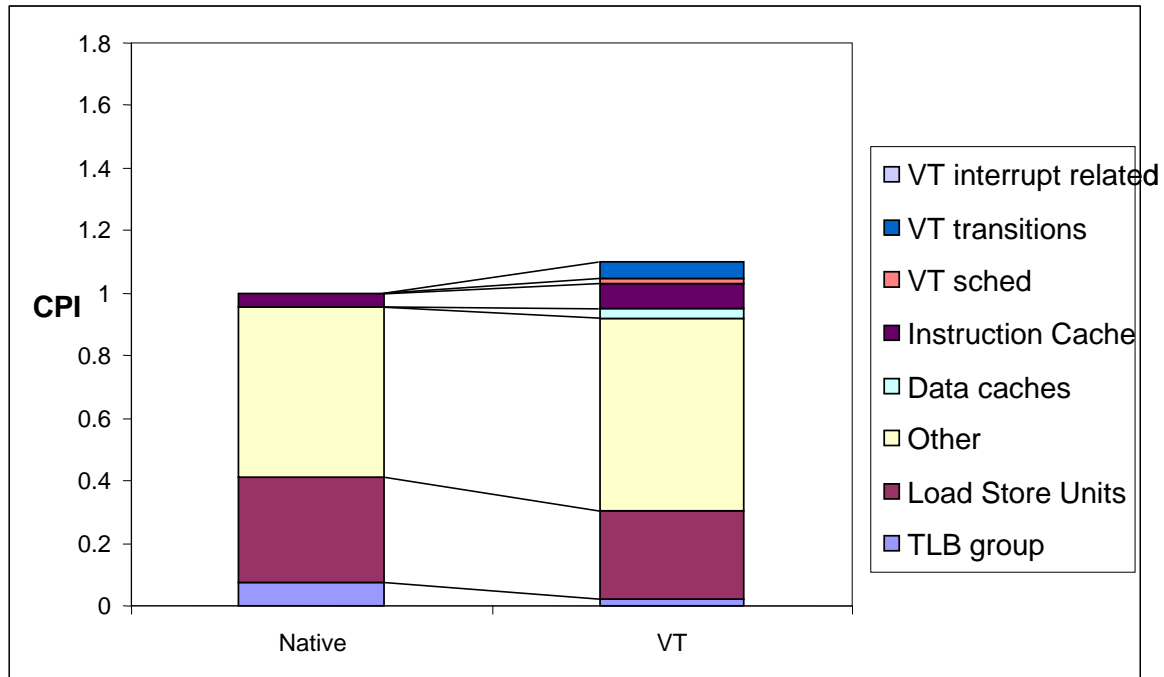


**Figure 2: CPI Inflation and Decomposition for Native and Virtualized Execution of Linux Build**

It is evident from the figure that for the Linux kernel build, virtualized execution generates a significant overhead. CPI for virtualized execution has an average value of 1.72, while the native case has an average CPI of only 1.03. Such large increase is understandable for this workload as the build activity is process creation intensive, and gives rise to large numbers of address space setup and teardown activities. This leads to many transitions to the hypervisor as a result of copy-on-write fault handling. Hardware assisted memory virtualization should remove the majority of such transitions in future processors.

The figure also indicates some of the key contributors to the architectural increase in CPI. It is clear that the load/store operations are having a much higher impact in the case of virtualized execution than in the native case. This is not surprising since virtualization can increase resource

pressures and result in these operations experiencing increased cache and TLB misses. Among the VT events, it is clear that VT transitions have a significant impact on the performance of this workload. In fact, more than 20% of the overall workload execution time (and close to half of the VT overhead) is attributable to this event.



**Figure 3: Inflation and Decomposition for Native and Virtualized Execution of Dacapo Suite**

In a similar way, Figure 3 provides the native and VT CPI decomposition for the Dacapo suite benchmark. For this workload, virtualization has a relatively low overhead: only 9%. The figure indicates that this overhead is mainly due to VT transitions and scheduling. We can also observe a change in the relative distribution of architectural contribution to the CPI, with an increase in pressure on the data caches, and an increase in the base CPI (the "Other" group).

The main drawback of multi-linear regression is that it can suffer from low accuracy since it cannot account for the non linear interaction between the independent variables. To assess the validity of the conclusions, we evaluate the different performance models using two main accuracy metrics:

- a. The correlation coefficient between the predicted and actual CPI value.
- b. The average absolute error: difference between actual and predicted CPI.

As shown in Table 5, the native and virtualizations models for both benchmarks give encouraging accuracy numbers. In the next sub-section, we introduce the use of Model trees for as mechanism to gain insight into the difference between performance under native and virtualized executions. We are currently investigating the usefulness of Model trees and other regression techniques to perform a detailed CPI decomposition.

This section demonstrated how simple multi-linear regression can be used to assess VT overhead to determine the most important micro-architectural and virtualization events causing this overhead for a particular workload. In the next subsection we apply a non-linear analysis technique, Model Trees, to evaluate the relative importance of various virtualization and architectural factors in predicting the CPI metric.

Model	Correlation Coefficient	Average Absolute Error
Linux Native	83%	6.0%
Linux Virtualization	88%	8.72%
Dacapo Native	75.6%	10.72%
Dacapo Virtualization	80.91	10.09%

**Table 6: Models Accuracy**

## 5.2 Model Tree analysis for VT CPI Estimation

In this section, we apply Model Tree analysis [19, 20] to assess the effects of virtualization on workload performance. This approach allows for handling non-linear interactions between the different independent variables. To apply Model trees, we merge the two data sets (native and VT execution). The VT events that are available for virtualized execution are replaced with a single “VT” variable, which is either 0 or 1, respectively to indicate whether a given data point comes from the native or virtualized execution. This modification was able to present a better convergence for the models. In future work we will investigate the optimal set of virtualization events that can produce a converging Model Tree with high accuracy.

There are two main components to the generated Model Trees: the Model Tree itself and a set of linear equations at the leaf nodes of the tree. Figure 4 presents the tree for the Linux build benchmark. In this tree, the first split event in the tree is indeed the VT variable (VT > 0.5 simply indicates that virtualization was enabled). This means that the factor that has the most effects on the CPI for this workload is whether this was a virtualized execution. The fact that the model clearly splits the data into distinct groups also indicates that performance models for native and virtualized executions are inherently different even when evaluated primarily using micro-architectural events. This empirically confirms our research motivation; namely of the need for building new performance and projection models for VT. In the Model Tree, the linear equations can be found at the leaf level nodes labeled as LM01, LM02 ... etc. For reasons of space, the linear model equations have been left out of this omitted and only the average CPI and the representation within the data sample are included for each linear grouping.

```

VT <= 0.5 [50%, MEAN CPI=1.032]
| ITLB_MISSES <= 0.000258 [34%, MEAN CPI=0.973]
| | DIV_OPS <= 0.000022 [10%, MEAN CPI=0.875]
| | | L2_MISSES <= 0.000034 [2%, MEAN CPI=0.724]----- LM01
| | | L2_MISSES > 0.000034 [8%, MEAN CPI=0.913]----- LM02
| | | DIV_OPS > 0.000022 [24%, MEAN CPI=1.015]----- LM03
| | ITLB_MISSES > 0.000258 [16%, MEAN CPI=1.157]----- LM04
VT > 0.5 [50%, MEAN CPI=1.723]
| DTLB_misses <= 0.0045855 [32%, MEAN CPI=1.454]----- LM05
| DTLB_misses > 0.0045855 [18%, MEAN CPI=2.196]
| | DIV_OPS <= 0.0005335 [5%, MEAN CPI=1.860]
| | | ITLB_MISSES <=0.002224 [3%, MEAN CPI=1.662]----- LM06
| | | ITLB_MISSES > 0.00222 [2%, MEAN CPI=2.142]----- LM07
| | | DIV_OPS > 0.0005335 [13%, MEAN CPI=2.331]
| | | L2_MISSES <= 0.0007365 [9%, MEAN CPI=2.230]
| | | | L2_MISSES<=0.000314 [2%, MEAN CPI=2.037]----- LM08
| | | | L2_MISSES> 0.000314 [7%, MEAN CPI=2.280]----- LM09
| | | L2_MISSES > 0.0007365 [4%, MEAN CPI=2.486]
| | | | Loads <= 0.24908 [1%, MEAN CPI=2.295]----- LM10
| | | | Loads > 0.24908 [3%, MEAN CPI=2.573]----- LM11

```

**Figure 4: Model Tree for the Linux Build**

```

L1I_MISSES <= 0.0004695 [41%, MEAN CPI=0.880]
| Store <= 0.1375 [11%, MEAN CPI=0.688] -----LM01
| Store > 0.1375 [30%, MEAN CPI=0.949]
| | L1I_MISSES <= 0.0000011 [7%, MEAN CPI=0.817]

```

```

DTLB_misses <= 0.00008445 [5%, MEAN CPI=0.768]
| BR_RETIREED <= 0.1765 [1%, MEAN CPI=0.623]----- LM02
| BR_RETIREED > 0.1765 [4%, MEAN CPI=0.808]
| | L1D_MISSES <= 0.00279 [2%, MEAN CPI=0.858]----- LM03
| | L1D_MISSES > 0.00279 [2%, MEAN CPI=0.752]----- LM04
DTLB_misses > 0.00008445 [2%, MEAN CPI=0.928]----- LM05
L1I_MISSES > 0.0000011 [23%, MEAN CPI=0.992]
DTLB_misses <= 0.0003075 [17%, MEAN CPI=0.962]
| BR_RETIREED <= 0.1885 [6%, MEAN CPI=1.017]
| | PAGE_WALKS <= 0.00035 [1%, MEAN CPI=0.858]----- LM06
| | PAGE_WALKS > 0.00035 [5%, MEAN CPI=1.056]
| | | L1D_MISSES <= 0.000426 [2%, MEAN CPI=1.113]
| | | | Loads <= 0.404 [1%, MEAN CPI=1.03]----- LM07
| | | | Loads > 0.404 [2%, MEAN CPI=1.165]----- LM08
| | | L1D_MISSES > 0.000426 [3%, MEAN CPI=1.001]
| | | | Loads <= 0.365 [1%, MEAN CPI=1.112]----- LM09
| | | | Loads > 0.365 [2%, MEAN CPI=0.92]----- LM10
| BR_RETIREED > 0.1885 [10%, MEAN CPI=0.929]
| | BR_MISPRED <= 0.00561 [4%, MEAN CPI=0.913]----- LM11
| | BR_MISPRED > 0.00561 [6%, MEAN CPI=0.939]
| | | MUL_OPS <= 0.0007005 [2%, MEAN CPI=0.884]----- LM12
| | | MUL_OPS > 0.0007005 [4%, MEAN CPI=0.972]----- LM13
DTLB_misses > 0.0003075 [6%, MEAN CPI=1.080]
| BR_RETIREED <= 0.2115 [5%, MEAN CPI=1.119]
| | Loads <= 0.3975 [4%, MEAN CPI=1.164]----- LM14
| | Loads > 0.3975 [1%, MEAN CPI=0.958]----- LM15
| BR_RETIREED > 0.2115 [1%, MEAN CPI=0.930]----- LM16
L1I_MISSES > 0.0004695 [59%, MEAN CPI=1.139]
PAGE_WALKS <= 0.002755 [43%, MEAN CPI=1.073]
| Store <= 0.2005 [13%, MEAN CPI=1.015]
| | MUL_OPS <= 0.0004145 [5%, MEAN CPI=0.874]----- LM17
| | MUL_OPS > 0.0004145 [8%, MEAN CPI=1.094]
| | | L1I_MISSES <= 0.001335 [2%, MEAN CPI=0.989]----- LM18
| | | L1I_MISSES > 0.001335 [6%, MEAN CPI=1.135]----- LM19
| Store > 0.2005 [30%, MEAN CPI=1.099]
| | L1I_MISSES <= 0.007505 [24%, MEAN CPI=1.073]
| | | BR_RETIREED <= 0.1875 [12%, MEAN CPI=1.105]
| | | | L1D_MISSES <= 0.002565 [9%, MEAN CPI=1.085]
| | | | | Loads <= 0.4105 [4%, MEAN CPI=1.042]----- LM20
| | | | | Loads > 0.4105 [5%, MEAN CPI=1.130]
| | | | | | L1D_MISSES <= 0.000359 [3%, MEAN CPI=1.162]- LM21
| | | | | | L1D_MISSES > 0.000359 [2%, MEAN CPI=1.083]- LM22
| | | | L1D_MISSES > 0.002565 [4%, MEAN CPI=1.152]
| | | | | BR_RETIREED <= 0.1595 [1%, MEAN CPI=1.31]----- LM23
| | | | | BR_RETIREED > 0.1595 [3%, MEAN CPI=1.088]----- LM24
| | | BR_RETIREED > 0.1875 [11%, MEAN CPI=1.039]
| | | | MUL_OPS <= 0.0009685 [4%, MEAN CPI=1.034]----- LM25
| | | | MUL_OPS > 0.0009685 [7%, MEAN CPI=1.042]----- LM26
| | L1I_MISSES > 0.007505 [6%, MEAN CPI=1.200]
| | | L2_MISSES <= 0.000306 [5%, MEAN CPI=1.163]
| | | | DIV_OPS <= 0.0001895 [3%, MEAN CPI=1.179]
| | | | | BR_RETIREED <= 0.1595 [1%, MEAN CPI=1.315] ----- LM27
| | | | | BR_RETIREED > 0.1595 [2%, MEAN CPI=1.142]
| | | | | | DIV_OPS <= 0.00001065 [1%, MEAN CPI=1.073]-- LM28
| | | | | | DIV_OPS > 0.00001065 [1%, MEAN CPI=1.16]--- LM29
| | | | | DIV_OPS > 0.0001895 [2%, MEAN CPI=1.135] ----- LM30
| | | | L2_MISSES > 0.000306 [1%, MEAN CPI=1.348]----- LM31
PAGE_WALKS > 0.002755 [16%, MEAN CPI=1.315]----- LM32

```

**Figure 5: Model Tree for the Dacapo Suite**

Another use of this Model tree analysis is to provide an alternative way of estimating the average VT overhead: this can easily be done by determining the difference in the average CPI between

the sub-tree on the left and the one on the right of the root node. Since the model only uses micro-architectural events, it can also be used to estimate how much of the overhead is attributable to the way virtualized execution exercises micro-architectural resources (explained using the variables in the sub-tree for virtualization) and how much can only be explained by including VT events. The later part could be viewed as dependent on VT architecture.

We also applied Model Trees to analyze to the Dacapo suite benchmark and resultant model is shown in Figure 5. Unlike the Linux benchmark, the VT dummy variable does not appear at the root of the tree. In fact, it does not show up as a split variable. Instead, it appears in all the linear equations with different coefficients in the different models. Like the Linux benchmark model, the Dacapo Model Tree indicates that virtualization is associated with a statistical degradation in performance: that is, the coefficient in front of the VT dummy variable is strictly positive in all the linear models.

## 6 Related Work

Several previous studies have examined specific impacts of virtualization on system performance. [2] studies the network processing slowdown due to virtualization with Xen, while [7] examines the overheads from I/O processing in the hypervisor due to requests from specific virtual machines. [16] analyzes the cross interference of concurrently executing guests and [26] evaluates the impact of Xen based paravirtualization on the message passing interface and process execution in communication intensive High-Performance-Computing (HPC) systems.

Current studies in virtualization often rely on execution sampling utilities or extensions to native systems. Xenoprof is a Xen profiling utility based on the popular Oprofile system used for native UNIX based systems [17]. Xenoprof extends the profiling capabilities of Oprofile by distributing measured performance counter data to the corresponding guest systems. Xenoprof, through extended Oprofile utilities, then assigns performance events to specific code regions from the kernel and the application. Perfctr and PAPI [5] are two additional utilities commonly used for performance counter measurements and correlations. Neither currently supports virtualized platforms. Xentrace [25], Xentop [9], and XenMon [10] provide instrumentation implemented with the Xen system for counting hypervisor events and for tracing of hypervisor. We utilize Xentrace to collect hypervisor events in addition to the events we collect from the hardware PMC and the guest kernels. In addition to the facilities provided by these utilities, we also collect a synchronized vertical profile of the execution stream including the entire software and hardware stack. In this respect our approach is similar to the Vertical Profiling system described in [11] for the Jikes Java Virtual Machine.

As hardware and software systems become increasingly complex, statistical techniques for performance modeling [13, 23, 14, 20, 19] have become invaluable design and optimization tools for gaining insights into performance critical factors. Several studies also propose performance models to reduce the number of simulations required to explore the design space of processors and other systems on chip. [23] presents a model for representing performance metrics as a linear function of micro-architectural events and uses it to improve the process of design space exploration for embedded processors. Recently more complex analytical models have been applied to benchmark performance analysis. [13] proposes a model based on Artificial Neural Networks. [14] uses a non-linear model based on radial basis function networks. [19] compares five machine-learning regression algorithms applied to PMU data for a subset of SPEC CPU2006.

Model trees were found to perform as well as artificial neural networks (ANNs) and support vector machines (SVMs) and had the advantage of interpretability. [20] describes the application of Model trees for charactering benchmark performance.

Finally, [6] proposed a benchmarking suite for evaluating and comparing Virtual Machine Monitors. This study is an important step towards standardizing the measurement criteria for comparing hypervisor performance. The methodology and instrumentation approach we have advanced in this paper can aid exploring in detail the performance of virtual machine monitors and to assess the architectural and system level impacts of virtualization under such a benchmarking suite.

## 7 Conclusion

We described an analysis approach that allows one to break down the cycles per instruction (CPI) performance metric for a workload using statistical regression. The specific challenge our approach addresses is that of obtaining the performance measurements at a fine granularity of sampling, such that factors of interest in software and hardware are counted in unison and at very low perturbation to the system. We demonstrated by example how such instrumentation allowed us to deconstruct the CPI metric for two different workloads, using two different regression algorithms: first, a multi-linear fit against performance data at the whole workload level, and next, a Model-tree characterization of the workloads using fine grained data gathered at the granularity of several million instructions. Our work is particularly useful in the context of execution in virtual machines because it removes the need to assume single state behaviors over long intervals of time in order to relate CPI to its factors.

This work is a first step on a path to building more complex and detailed predictive models for workload performance. Our next steps are to extend the instrumentation and the methodology so that resource utilization under homogenous and heterogeneous consolidation of workloads can be measured accurately and used in statistical prediction of overall workload performance.

## References

1. Adams K. and Agesen O. A comparison of software and hardware techniques for x86 virtualization. In *Proceedings of the 12th international Conference on Architectural Support for Programming Languages and Operating Systems*. Oct 2006.
2. Apparao P., Makineni S., and Newell D. Characterization of network processing overheads in Xen. In *Proceedings of the 2nd international Workshop on Virtualization Technology in Distributed Computing*. Nov 2006.
3. Barham P., Dragovic B., Fraser K., Hand S., Harris T., Ho A., Neugebauer R., Pratt I., and Warfield A. 2003. Xen and the art of virtualization. In *Proceedings of the Nineteenth ACM Symposium on Operating Systems Principles*. Oct 2003.
4. Blackburn, S. M., Garner, R., Hoffman, C., Khan, A. M., McKinley, K. S., Bentzur, R., Diwan, A., Feinberg, D., Frampton, D., Guyer, S. Z., Hirzel, M., Hosking, A., Jump, M., Lee, H., Moss, J. E. B., Phansalkar, A., Stefanovic, D., VanDrunen, T., von Dincklage, D., and Wiedermann, B. The DaCapo Benchmarks: Java Benchmarking Development and Analysis. In *Proceedings of the 21st annual ACM SIGPLAN conference on Object-Oriented Programming, Systems, Languages, and Applications*. Oct, 2006.

5. Browne S., Dongarra, J., Garner N., Ho G., Mucci P. A Portable Programming Interface for Performance Evaluation on Modern Processors. In *The International Journal of High Performance Computing Applications*, Volume 14, number 3, 189-204. Fall 2000.
6. Casazza, J., Greenfield, M., Shi, K. Redefining Server Performance Characterization for Virtualization Benchmarking. In *Intel® Technology Journal*. Aug 2006.
7. Cherkasova L. and Gardner R. 2005. Measuring CPU overhead for I/O processing in the Xen virtual machine monitor. In *Proceedings of the USENIX Annual Technical Conference*, 24-24. Apr 2005
8. Eyerman S., Eeckhout L., Karkhanis T., and Smith J. E. 2006. A performance counter architecture for computing accurate CPI components. In *Proceedings of the 12th international Conference on Architectural Support for Programming Languages and Operating Systems*. Oct 2006.
9. Fischbach J., Hendricks D., and Triplett J. Xentop (Xen utility). 2005.
10. Gupta D., Gardner R., Cherkasova L. XenMon: QoS Monitoring and Performance Profiling Tool. *HP Technical Report*, HPL-2005-187. Oct 2005.
11. Hauswirth M., Diwan A., Sweeney P. F., and Mozer M. C. 2005. Automating vertical profiling. In *Proceedings of the 20th Annual ACM SIGPLAN Conference on Object Oriented Programming, Systems, Languages, and Applications*, 281-296. Oct 2005.
12. Intel® Corporation. Intel64 and IA-32 Architectures Software Developers Manual. 2006
13. İpek, E., McKee, S. A., Caruana, R., de Supinski, B. R., and Schulz, M. Efficiently exploring architectural design spaces via predictive modeling. In *SIGOPS Operating System Review*. 195-206. Oct. 2006.
14. Joseph P. J., Vaswani K., and Thazhuthaveetil M. J. A Predictive Performance Model for Superscalar Processors. In *Proceedings of the 39th Annual IEEE/ACM international Symposium on Microarchitecture*, 161-170. Dec 2006.
15. Lee B. C. and Brooks D. M. Accurate and efficient regression modeling for microarchitectural performance and power prediction. In *Proceedings of the 12th international Conference on Architectural Support for Programming Languages and Operating Systems*. Oct 2006.
16. Matthews J. N., Hu W., Hapuarachchi M., Deshane T., Dimatos D., Hamilton G., McCabe M., and Owens J. Quantifying the performance isolation properties of virtualization systems. In *Proceedings of the 2007 Workshop on Experimental Computer Science*. Jun 2007.
17. Menon A., Santos J. R., Turner Y., Janakiraman G., Zwaenepoel W. Diagnosing Performance Overheads in the Xen Virtual Machine Environment, In *First ACM/Usenix Conference on Virtual Execution Environments*. Jun 2005.
18. Novell Inc., "Virtualization in the Data Center", Technical White Paper, #462-002015-001. May 2006.
19. Ould-Ahmed-Vall E., Woodlee J., Yount C., and Doshi K.A. On the Comparison of Regression Algorithms for Computer Architecture Performance Analysis of Software Applications. In *First Workshop on Statistical and Machine learning approaches applied to ARchitectures and compilation*. Jan 2007.
20. Ould-Ahmed-Vall E., Woodlee J., Yount C., Doshi K.A., Abraham S. Using Model Trees for Computer Architecture Performance Analysis of Software Applications. In *Performance Analysis of Systems & Software*, 116-125. Apr 2007.
21. Rosenblum M., Garfinkel T. Virtual machine monitors: current technology and future trends, *Computer*, 39-47. May 2005.
22. Sherwood S., Sair S., and Calder B. Phase Tracking and Prediction, In *Proceedings of the 30th Annual Intl. Symposium on Computer Architecture*. Jun 2003.
23. Simonson L. J. and He L. Micro-Architecture Performance Estimation by Formula, In *Embedded Computer Systems: Architectures, Modeling, and Simulation*, 192-201. Jul 2005

24. Smith L. InformationWeek survey shows virtualization taking off in the data center. In *Information Week*, Feb 10, 2007.
25. Williamson M. Xentrace (Xen utility).
26. Youseff L., Wolski R., Gorda B., and Krintz C. Evaluating the Performance Impact of Xen on MPI and Process Execution For HPC Systems. In *Proceedings of the 2nd international Workshop on Virtualization Technology in Distributed Computing*. Nov 2006.