Performance Metrics for Consolidated Servers

Andy Georges Lieven Eeckhout
Ghent University, Belgium
{andy.georges, lieven.eeckhout}@elis.ugent.be

Abstract

In spite of the widespread adoption of virtualization and consolidation, there exists no consensus with respect to how to benchmark consolidated servers that run multiple guest VMs on the same physical hardware. For example, VMware proposes VMmark which basically computes the geometric mean of normalized throughput values across the VMs; Intel uses vConsolidate which reports a weighted arithmetic average of normalized throughput values.

These benchmarking methodologies focus on total system throughput (i.e., across all VMs in the system), and do not take into account per-VM performance. We argue that a benchmarking methodology for consolidated servers should quantify both total system throughput and per-VM performance in order to provide a meaningful and precise performance characterization. We therefore present two performance metrics, Total Normalized Throughput (TNT) to characterize total system performance, and Average Normalized Reduced Throughput (ANRT) to characterize per-VM performance.

We compare TNT and ANRT against VMmark using published performance numbers, and report several cases for which the VMmark score is misleading. This is, VMmark says one platform yields better performance than another, however, TNT and ANRT show that both platforms represent different trade-offs in total system throughput versus per-VM performance. Or, even worse, in a couple cases we observe that VMmark yields opposite conclusions than TNT and ANRT, i.e., VMmark says one system performs better than another one which is contradicted by the TNT/ANRT performance characterization.

Categories and Subject Descriptors D4.8 [*Operating systems*]: Performance—Measurements

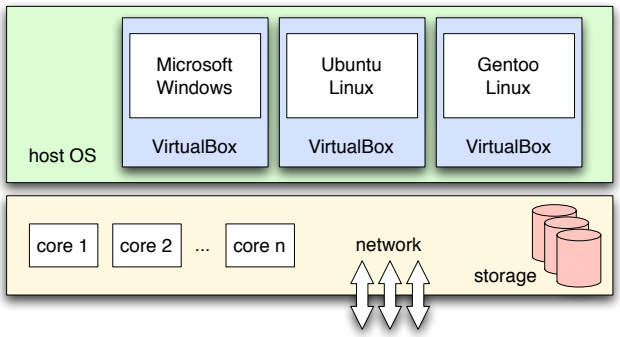
General Terms Experimentation, Measurement, Performance

Keywords virtualization, consolidation, benchmarking, performance, metrics

1. Introduction

Over the past few years, system virtualization has gained renewed interest. Trends towards multi-core processing have led to the proliferation of relatively inexpensive and powerful processors, yet, applications do not fully utilize these systems. System virtualization allows for running multiple virtual machines (VMs) on the same physical hardware — called server consolidation — thereby increasing system utilization and reducing cost. Although the physical hardware is shared across the virtual machines, the virtual machines are fully isolated from each other, and each virtual machine runs a separate operating system (OS) instance and application software. Server consolidation is becoming commonplace on a wide range of systems, from desktop computers to heavy-duty server blades. For example, on the desktop, one can employ VMWare's Workstation or Sun's VirtualBox to run applications under Microsoft Windows on top of a Linux host OS. On the server side, one can have container-based VMs (e.g., OpenVZ, FreeBSD jails) that each run a separate instance of the same OS. Alternatively, using a bare-metal virtual machine monitor (VMM), such as VMWare ESX or Xen, the server can run a heterogeneous mix of OSes and application stacks simultaneously. Figure 1 illustrates a few of these potential setups.

(a) Desktop virtualization



(b) Server virtualization

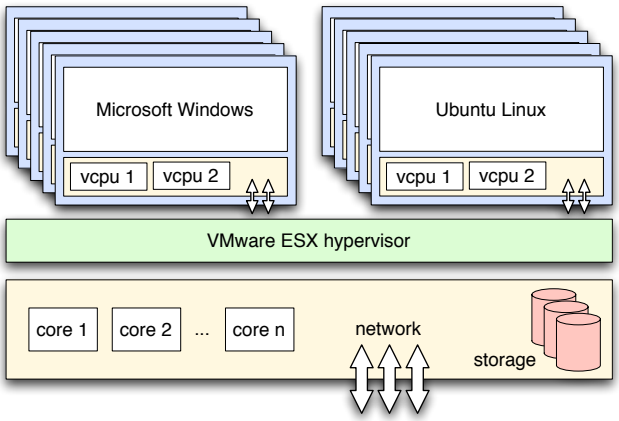


Figure 1: Two typical examples of consolidated setups, using (a) desktop virtualization with Sun's VirtualBox, and (b) server virtualization with VMware's ESX.

HPCVirt'10 April 13, 2010, Paris.

It is to be expected that this trend towards server consolidation is going to continue in the coming years, for two reasons. For one, we have entered the multi-core era: contemporary processors integrate multiple processor cores on a single chip, and given Moore's law which predicts exponential growth in transistor density with each technology generation, increasing core counts are likely. Moreover, simultaneous multi-threading (SMT), or running multiple hardware threads on a single core is commonplace in today's commercial processors, e.g., Intel Core i7, IBM Power 6, etc. As such, very soon there will be multiple tens or even hundreds of hardware threads running on a single processor — in fact, some contemporary processors already run tens of hardware threads. Second, virtualization overhead is steadily decreasing thanks to advances in hardware support, e.g., Intel's VT-x includes Extended Page Tables (EPT), and AMD-V provides Nested Page Tables. Decreased virtualization overhead allows for greater consolidation ratios, i.e., the number of system virtual machines that can be run on a single physical machine keeps increasing.

In spite of the long history in virtualization and its renewed interest for server consolidation, there is no consensus on how to benchmark consolidated servers. However, it is paramount that in order to make meaningful performance comparisons across systems one needs a rigorous benchmarking methodology. A benchmarking methodology involves a number of dimensions, such as its experimental setup, its benchmarks and its metrics. In this paper, we focus on the latter: without meaningful and precise performance metrics it is hard to make an assessment about the performance differences seen across systems. Imprecise metrics can potentially lead to incorrect, or at least misleading, conclusions.

Even when it comes to metrics there is no consensus about what is the right performance metric for consolidated servers. VMware's VMmark benchmarking methodology [8] considers the geometric mean of normalized throughput values across the VMs in a tile; each tile runs 5 non-idle VMs and one idle VM, and the overall VMmark score is the sum of the geometric means across all tiles. VMmark reports this overall score along with the number of tiles supported. Intel's vConsolidate [1, 2] computes the weighted arithmetic average of the normalized throughput values across all VMs consolidated on the server.

A major limitation for existing benchmarking methodologies is that they report a single performance number that primarily focuses on total aggregate system performance. We conjecture that a benchmarking methodology for consolidated servers should not only focus on total system throughput but should also consider per-VM performance. The pitfall in focusing on total system throughput is that the performance metric is biased towards systems that prioritize easy-to-virtualize applications over hard-to-virtualize applications. This is, prioritizing easy-to-virtualize applications may yield artificially high performance numbers, while some hard-to-virtualize applications observe severe performance penalties and may even starve. Current benchmarking approaches do not explicitly quantify per-VM performance. They rather capture per-VM performance in an ad-hoc way, or report a unified performance number that captures both total system performance and per-VM performance. For example, VMmark reports both total performance and the number of tiles; if per-VM performance decreases with increased tile counts, this will be reflected in the overall score. Or, one reports throughput numbers with quality-of-service (QoS), i.e., a transaction is only counted in the overall system throughput metric if it meets the QoS requirements [6].

In this paper, we propose two novel performance metrics for consolidated servers: Total Normalized Throughput (TNT) and Average Normalized Reduced Throughput (ANRT). Both metrics have a system-level meaning. TNT is a system-oriented metric and quantifies total aggregate system throughput; ANRT is a VM-oriented metric and quantifies the reduction in per-VM performance due to consolidation. We advocate that a performance study should report both metrics when benchmarking consolidated servers: they both characterize a different aspect of the perceived system performance. From a system's perspective, one typically cares about optimizing utilization and total system throughput, whereas end user concerns shift the focus towards per-VM performance.

In order to demonstrate the value of TNT versus ANRT performance characterization, we use publicly available performance data for a range of commercial systems and compare against VMmark. We show that VMmark performance characterization can be misleading in some cases. This is, VMmark says one platform yields better performance than another, however, TNT and ANRT show that both platforms represent different trade-offs in terms of total system throughput versus per-VM performance. Or, even worse, in a couple cases we observe that VMmark yields opposite conclusions than TNT and ANRT, i.e., VMmark says system A is better than system B, whereas a TNT/ANRT performance characterization shows that system B is better than system A.

We believe that a discussion on performance metrics for consolidated servers is timely. As mentioned before, there exist different approaches today and there is no consensus about what is the right approach. In addition, the SPEC consortium has installed a subcommittee to standardize virtualized server benchmarking. We hope this work will help find consensus about how to report consolidated server performance: we advocate that performance should be measured along two angles considering total system throughput and per-VM performance.

2. Existing benchmarking methodologies

There exist two well-known benchmarking methodologies for consolidated servers, namely VMware's VMmark [8] and Intel's vConsolidate [2]. We now revisit these initiatives.

2.1 VMmark

VMmark [8] is the virtualization benchmarking framework developed by VMware. Its basic unit of work is a so-called *tile*, which consists of 6 VMs, including one idle VM and 5 non-idle VMs. Each non-idle VM runs a guest OS with a benchmark application on top of it. VMmark considers the following benchmarks and each of these benchmarks runs in a separate VM: a mail server (Microsoft Exchange 2003), a Java server (SPECjbb2005), a web server (SPECweb2005), a database server (Oracle Swingbench), and a file server (dbench). Two benchmarks — the mail server and the Java servers — run on top of Microsoft Windows Server 2003. The others run on top of Suse Linux Enterprise Edition 10. The framework assigns 2 vCPUs (virtual CPU) and 2GB RAM to each VM; except for the VM running dbench, which gets allocated only one vCPU along with 256MB RAM.

The idea behind the notion of a tile is that the total workload of the consolidated server is increased by adding tiles, with the additional effect that each tile in the larger workload might measure lower performance. However, the total aggregate performance of the consolidated system should be larger, if the system is not overcommitted. VMmark reports a performance score that quantifies aggregate performance along with the number of tiles supported by the system.

The VMmark performance score is computed as follows, see also Figure 2. After an initial warm-up period, the performance (throughput) for each benchmark in the various tiles is measured during three consecutive 40-minute intervals. Each of these values is normalized with respect to the corresponding reference measurements. (The reference throughput values have been computed on a reference machine.) The geometric mean is then computed across the normalized throughput numbers within a tile — this is the geometric mean of 5 normalized throughput values, one throughput value per non-idle VM. This is done for all three 40-minute in-

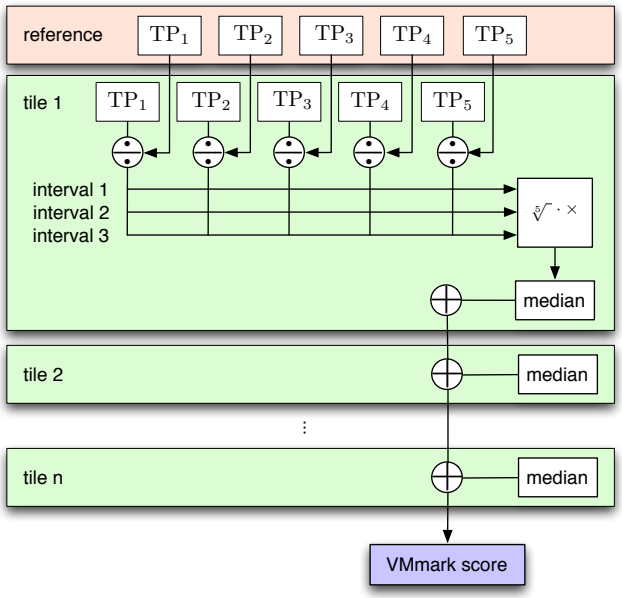


Figure 2: The scoring methodology used by VMmark for benchmarking consolidated servers.

tervals, hence there are three scores per tile. The median score is the score of the given tile. Finally, the median values are summed across all the tiles to obtain the final score.

2.2 vConsolidate

Intel's vConsolidate [1, 2] benchmarking methodology takes another approach and considers multiple VMs, with each VM running a benchmark. Example benchmarks are a web server, e-mail server or database server. The system can have multiple VMs running the same benchmark. Each benchmark has a designated weight; this weight is fixed and is pre-defined as part of the consolidation workload. For each benchmark, a reference performance number is determined, e.g., through a baseline measurement without virtualization.

The consolidation workload then consists of multiple (replicated) VMs along with an idle VM. vConsolidate computes the ratio of the performance under consolidation versus the reference run for each VM; this is a normalized throughput number. The overall performance score then is the weighted arithmetic sum across the normalized throughput numbers for the different (non-idle) VMs.

2.3 SPECvirt

The SPEC consortium has also formed a subcommittee consisting of a broad range of companies such as AMD, Intel, IBM, HP, VMware, Microsoft, and many others. The goal of this subcommittee is to develop standardized methods for benchmarking 'virtualized performance for data center servers'. One of the topics of research is 'to provide a means to fairly compare server performance when running a number of virtual machines'. Unfortunately, there is no further information available at this point in time.

2.4 Benchmarking multi-threaded and multi-core hardware

The computer architecture community is facing similar concerns with respect to how to benchmark multi-core and multi-threaded hardware performance when co-executing multiple independent programs (called a multi-program workload). For some time, architects used IPC throughput as the overall system performance metric. However, IPC throughput has the detrimental effect that it favors processor architectures that prioritize high-IPC programs at the expense of low-IPC programs. As such, IPC throughput is largely abandoned. Besides IPC throughput, a number of other performance metrics exist: Snavely and Tullsen [9] proposed weighted average speedup, and Luo et al. [7] came up with the hmean performance metric. There is no consensus though on what metric to use, and in addition, there was no clear understanding of the system-level meaning for each of these metrics. Eyerman and Eeckhout [4] took a top-down approach and reasoned about how to benchmark multi-program workload performance starting from system-level performance concerns. They came up with two metrics, namely system throughput (STP), a system-oriented metric, and average normalized turnaround time (ANTT), a user-oriented metric. This work inspired us to come up with performance metrics for benchmarking consolidated servers. The key difference though is that the architecture community focuses on execution time per unit of work, however, for the workloads typically run on consolidated servers, the focus is on throughput, which requires a different perspective. Yet the philosophy is similar, i.e., we conjecture one needs to focus on both total aggregate system performance as well as user-perceived performance. The contribution of this paper is to translate this idea from the architecture community to benchmarking consolidated servers, and to contrast this new benchmarking approach against current practice.

3. TNT and ANRT

The workloads consolidated on servers are typically throughput-oriented. Hence, the basis for the TNT and ANRT performance metrics is per-VM throughput. In particular, for the different benchmarks in VMmark this is actions per minute for the mail server, orders per second for the Java server, commits per second for the database server, accesses per second for the web server, and megabytes per second for the file server. As in VMmark, we assume that each of these benchmarks runs in a separate VM.

As argued in the introduction, one should not only focus on total aggregate system throughput when benchmarking consolidated servers. Instead, one should also consider per-VM performance. The reason is that focusing on aggregate throughput only may lead to favoring systems that prioritize easy to virtualize and consolidate benchmarks, which may lead to unfair treatment and potentially starvation of hard to virtualize workloads. This insight has led us to propose two novel performance metrics for consolidated servers, namely TNT and ANRT, which we discuss in detail now.

3.1 Total Normalized Throughput

The goal of the Total Normalized Throughput (TNT) metric is to quantify total aggregate system performance. This is a system-oriented performance metric, and is of primary interest if one is interested in maximizing aggregate server or datacenter utilization and throughput.

To this end, we first define Normalized Throughput (NTP) as follows:

$$\mathrm{NTP}_i = \frac{\mathrm{TP}_i(\mathrm{V})}{\mathrm{TP}_i(\mathrm{R})}, \tag{1}$$

with $\mathrm{TP}_i(\mathrm{V})$ and $\mathrm{TP}_i(\mathrm{R})$ the throughput scores for VM i on the virtualized/consolidated system and on the reference platform, respectively. Normalization yields a dimension-less value which enables comparing relative performance across the different benchmarks in the consolidated workload.

Given the per-benchmark/VM normalized throughput NTP_i values, we now define Total Normalized Throughput (TNT):

$$\mathrm{TNT} = \sum_{i=1}^{n} \mathrm{NTP}_i. \tag{2}$$

TNT quantifies the total aggregate normalized throughput of the consolidated machine, i.e., it is a higher-is-better metric.

VMmark computes the geometric average across the normalized throughput scores within a tile. The controversy about how to compute average metrics across multiple benchmarks was comprehensively summarized by John [5]. She concluded that the geometric mean has no physical meaning. Essentially, the geometric mean is to be used whenever the aggregate value can be computed as the product of the individual measurements, which is not the case when determining the aggregate throughput in a consolidated system. Therefore, we advocate summing the normalized throughput values rather taking the geometric average. The intuition is that TNT quantifies the accumulated normalized throughput under consolidation.

3.2 Average Normalized Reduced Throughput

Increasing the number of VMs running on the consolidated server yields better aggregate system throughput. However, it may lead to reduced per-VM performance. The Average Normalized Reduced Throughput (ANRT) metric aims at quantifying the loss in per-VM performance due to consolidation.

We define Normalized Reduced Throughput (NRT) for benchmark/VM i as follows:

$$\mathrm{NRT}_i = \frac{\mathrm{TP}_i(\mathrm{R})}{\mathrm{TP}_i(V)} = \frac{1}{NTP_i}. \quad (3)$$

NRT indicates by how much throughput is reduced for each VM due to consolidation.

The Average Normalized Reduced Throughput (ANRT) is then defined as

$$\mathrm{ANRT} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{NRT}_i, \quad (4)$$

and quantifies the average normalized reduced throughput across all VMs. ANRT is a lower-is-better performance metric. The rationale for taking an arithmetic average across the NRT values is twofold (again, based on the insights by John [5] on how to compute average performance scores). First, we disregard the geometric average because it has no physical meaning, as mentioned before. Second, throughput is inversely proportional to response time. Hence, an alternative interpretation of the NRT metric is to say that it quantifies the increase in response time per benchmark/VM due to consolidation. Given that the response time of the reference machine then appears as the denominator in the NRT formula, we advocate taking the arithmetic average across the NRT values for computing ANRT. ANRT could thus be viewed as a measure for the average reduction in response time — this is consistent with the view that ANRT should be a user/VM-oriented performance metric.

3.3 Pareto frontier

Having computed both the system-oriented TNT metric as well as the VM-oriented ANRT metric, one can analyze performance of a consolidated server along these two complementary perspectives. Figure 3 gives an illustrative example of how one could analyze consolidated server performance in terms of ANRT versus TNT — the evaluation section in this paper will show similar graphs using real data. The vertical axis shows the reciprocal of ANRT, and the horizontal axis shows TNT. Servers with a high 1/ANRT and a high TNT appear in the upper right corner and deliver both high aggregate system throughput and high per-VM performance. However, it is not always possible to optimize for both metrics. As such, there exists a Pareto frontier that groups all the servers that achieve the best possible trade-off in system throughput versus per-VM performance. The requirement for a server to be a Pareto-optimal server is that there exists no other server that performs better on both performance metrics. Hence, for two servers appearing on the Pareto frontier, one cannot say that one server yields better consolidation performance than another. The performance difference really is a trade-off in system throughput versus per-VM performance, i.e., performance is better only along one dimension.

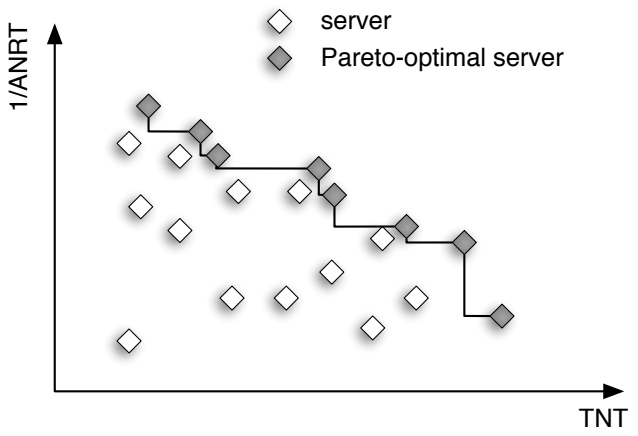


Figure 3: An illustrative example Pareto frontier in terms of TNT versus 1/ANRT.

4. Results

For evaluating the TNT and ANRT performance metrics, we use data that is publicly available on the VMmark website[1] as of October 14, 2009. We consider all the systems with 48, 32, 24 and 16 cores; there are 51 systems in total. For each of these systems, we compute the score for both the TNT and the ANRT metrics. TNT is computed as the sum of the normalized throughput values across all VMs and across all the tiles; ANRT is computed as the arithmetic average of the reciprocal of the normalized throughput values across all VMs and tiles. The TNT and ANRT scores are shown in Table 1 along with the number of machine cores, the date the experiment was conducted, the system vendor and model name/number, as well as the reported VMmark score and the number of tiles. The table has been sorted by VMmark score, i.e., systems with a high VMmark score are ranked higher — the rank number is shown in the leftmost column.

We organize the discussion in this evaluation section along four scenarios in which the TNT/ANRT performance characterization disagrees with the VMmark score. Each scenario reflects a different degree to which the metrics disagree. Each scenario will be illustrated in the following subsections by one or more case studies by comparing the systems listed in Table 1. These are the four scenarios:

- VMmark says one system is better than another, whereas TNT/ANRT characterization says the two systems represent different trade-offs in total system throughput versus per-VM performance.
- The VMmark score indicates that system A outperforms system B, whereas both TNT and ANRT indicate that B is better than A. This scenario shows strong disagreement between the VMmark and TNT/ANRT performance scores. There are a couple case studies that can be identified from Table 1 that fall under this scenario.
- The VMmark score indicates that system A outperforms system B, which is agreed upon by the TNT score, yet the ANRT score indicates the converse. A fairly large number of case studies can be identified that fall in this scenario.

[1] http://www.vmware.com/products/vmmark/results.html

rank	cores	experiment date	vendor	system	VMmark score	tiles	TNT	1/ANRT	ANRT
1	48	2009-08-24	HP	DL785	53.73	35	283.737	1.490	0.671
2	48	2009-08-11	HP	DL785g6	47.77	30	243.394	1.550	0.645
3	48	2009-07-28	NEC	5800 A1160	34.05	24	172.852	1.394	0.717
4	48	2009-06-16	IBM	X3950M2	33.85	24	172.818	1.380	0.725
5	32	2009-06-02	HP	DL785g5	31.56	21	160.793	1.472	0.679
6	32	2009-05-19	Unisys	7405R	30.86	20	164.650	1.502	0.666
7	32	2009-04-21	HP	DL785	30.5	21	147.766	1.432	0.698
8	24	2009-07-14	HP	DL585g6	29.95	20	151.922	1.468	0.681
9	24	2009-07-28	Dell	R905	29.51	20	149.654	1.446	0.692
10	32	2009-04-07	Unisys	7405R	29.19	19	148.729	1.501	0.666
11	24	2009-07-14	HP	BL685cg6	29.19	20	148.154	1.436	0.696
12	32	2009-03-10	Sun	X4600M2	29.11	19	147.786	1.496	0.668
13	32	2009-02-24	Unisys	7405R	28.97	19	147.523	1.490	0.671
14	32	2008-12-19	HP	DL785	27.71	19	140.399	1.431	0.699
15	32	2008-10-02	IBM	X3950	24.62	18	124.124	1.354	0.739
16	16	2009-06-18	Dell	M905	22.9	17	116.144	1.330	0.752
17	16	2009-05-19	Dell	R905	22.7	16	115.031	1.395	0.717
18	16	2009-04-24	HP	DL585G5	22.11	15	112.090	1.450	0.689
19	32	2008-11-05	Unisys	ES7000	21.96	15	111.453	1.434	0.697
20	32	2008-08-18	HP	DL785G5	21.88	16	110.944	1.349	0.741
21	16	2009-04-24	HP	BL685G6	20.87	14	105.495	1.462	0.684
22	24	2009-03-24	IBM	X3850M2	20.5	14	103.823	1.438	0.696
23	16	2009-01-27	HP	DL585G5	20.43	14	103.191	1.432	0.699
24	24	2009-02-24	IBM	X3850M2	20.41	14	103.371	1.431	0.698
25	16	2008-11-12	Dell	R905	20.35	14	103.124	1.425	0.702
26	24	2009-02-10	Dell	R900	19.99	14	101.186	1.403	0.713
27	16	2008-12-09	HP	BL685c	19.96	14	100.848	1.402	0.713
28	16	2008-11-12	Dell	M905	19.91	14	100.998	1.396	0.717
29	24	2009-01-27	Inspur	NF520D2	19.67	14	99.639	1.380	0.725
30	24	2009-01-13	Sun	X4450	19.47	14	98.546	1.375	0.727
31	16	2009-01-13	IBM	BladeLS42	19.17	14	96.744	1.355	0.738
32	24	2008-08-14	IBM	3850M2	19.1	14	96.583	1.353	0.739
33	24	2008-12-02	Dell	R900	18.69	14	94.917	1.323	0.756
34	24	2009-03-30	HP	BL680cG5	18.64	14	93.995	1.319	0.758
35	24	2008-10-06	HP	DL580G5	18.56	14	93.455	1.314	0.761
36	24	2008-07-08	Dell	R900	18.49	14	92.770	1.306	0.766
37	16	2008-11-12	IBM	BladeLS42	16.81	11	84.877	1.481	0.675
38	24	2008-09-10	HP	BL680cG5	16.05	12	81.048	1.323	0.756
39	16	2008-10-08	Dell	R905	15.35	11	77.792	1.371	0.729
40	16	2008-10-01	Dell	M905	15.09	11	76.549	1.355	0.738
41	16	2008-09-17	Dell	R905	14.84	10	75.132	1.448	0.691
42	16	2008-08-05	HP	DL585G5	14.74	10	82.061	1.443	0.693
43	16	2008-08-12	Dell	M905	14.28	11	72.038	1.285	0.778
44	16	2008-05-06	Dell	R905	14.17	10	72.076	1.393	0.718
45	16	2008-05-09	HP	DL580G5	14.14	10	71.159	1.394	0.718
46	16	2008-08-07	Dell	R900	14.05	10	70.923	1.386	0.722
47	16	2008-03-26	IBM	X3850M2	13.16	9	66.635	1.441	0.694
48	16	2008-04-25	SUN	X4450	12.23	8	62.040	1.489	0.672
49	16	2007-11-19	Dell	R900	12.23	8	62.048	1.487	0.672
50	16	2007-08-31	HP	DL580G5	11.54	8	58.335	1.413	0.708
51	16	2007-08-31	HP	BL680G5	10.17	7	51.549	1.426	0.702

Table 1: Public VMmark data ordered according to a descending VMmark score, extended with the TNT and ANRT metrics.

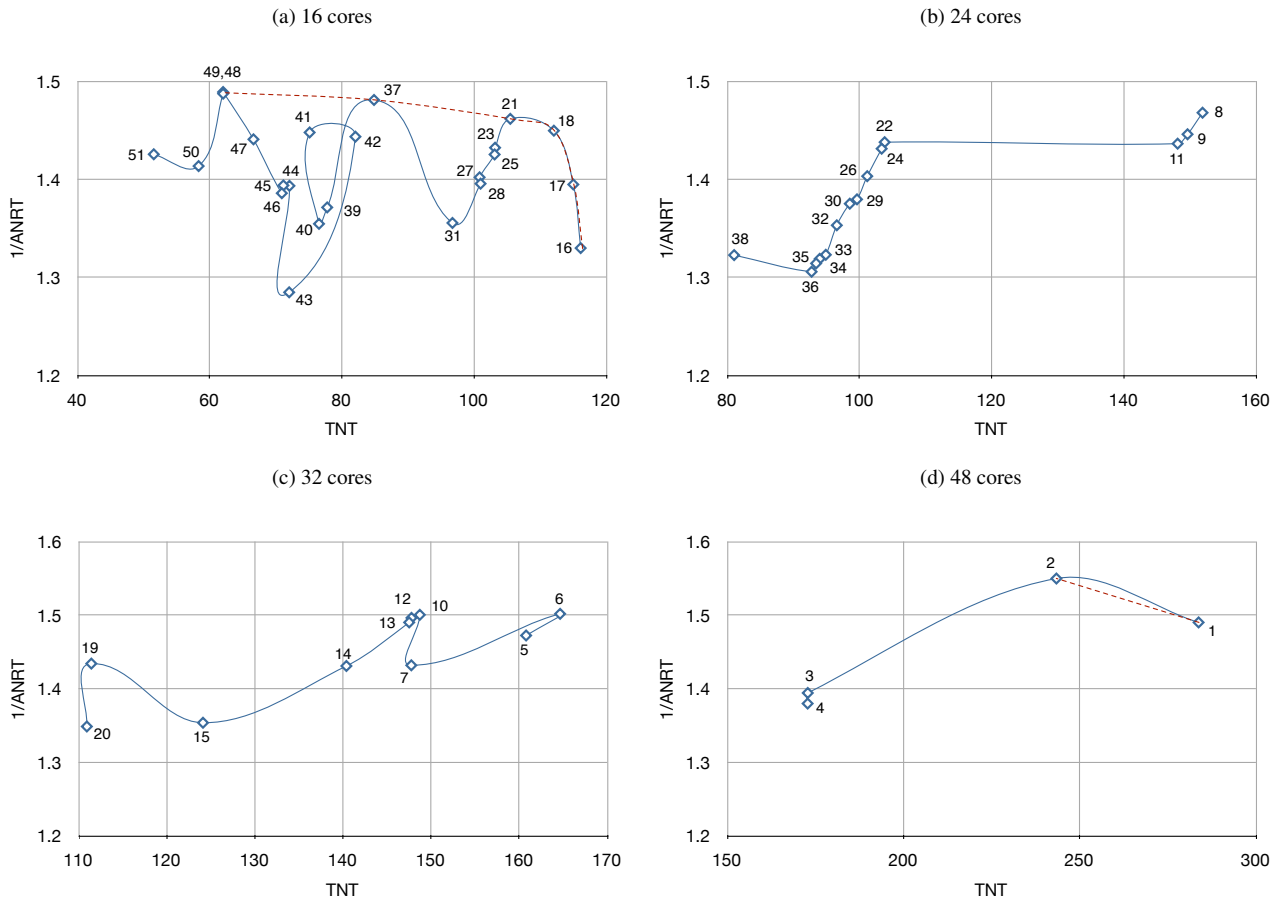


Figure 4: This figure shows graphs that set out the various systems from Table 1 in terms of 1/ANRT vs. TNT; there are separate graphs for separate core counts: (a) 16 cores, (b) 24 cores, (c) 32 cores, and (d) 48 cores. The solid line indicates the ordering of the systems according to the VMmark score. The dashed line shows the Pareto front, i.e., it connects the Pareto-optimal points in the graph. The numbers next to each point indicate the system's rank in the full set.

- The fourth scenario represents the case where the VMmark score indicates that system A outperforms system B, which is agreed upon by the ANRT score, yet the TNT score indicates the converse. There is one case study that falls under this scenario.

We now discuss each of the above scenarios in more detail. We will refer to each of the systems by their rank number as listed in the leftmost column in Table 1. During the discussion, we use the graphs shown in Figure 4. These graphs show the systems in terms of TNT versus 1/ANRT; there are separate graphs for the different core counts: (a) 16 cores, (b) 24 cores, (c) 32 cores, and (d) 48 cores. Each dot represents a system. The solid line connecting the various dots in each graph indicates the ordering of the systems according to the VMmark score; within each graph the lowest ranked system appears at the leftmost end of the line.

4.1 Scenario #1: TNT/ANRT Pareto frontier

The first scenario considers the case where VMmark says one system is better than another, whereas TNT versus ANRT performance characterization shows that there really is a trade-off in aggregate system throughput versus per-VM performance. Figures 4 (a) and (d) clearly show this trade-off in TNT and ANRT through the multi-point Pareto frontier. Essentially, a Pareto frontier explicitly states the trade-off in TNT versus ANRT: there is no other system that does better on both TNT and ANRT.

The frontier for the 16-core systems is formed by systems with rank numbers #48, #49, #37, #21, #18, #17, and #16. These systems outperform the other 16-core systems on both metrics, i.e., there is no system that yields a better score on either metric. On the other hand, we cannot readily state which Pareto-optimal system is better than any other Pareto-optimal system. If one values per-VM throughput more, one may be inclined to choose the system with the highest 1/ANRT score, i.e., system with rank #48. However, at a small sacrifice in ANRT, it is possible to gain significantly in aggregate system throughput: then system ranked #21 is an interesting design point. If one is willing to sacrifice an additional 9% increase in ANRT, one can gain up to 3.5% in TNT by selecting system ranked #16.

A similar reasoning is possible for the 48-core systems. Here, the frontier is smaller — partly because of the smaller number of systems in total. The Pareto frontier consists of systems with ranks #1 and #2. The trade-off here is a 16% gain in TNT versus a 4% reduction in ANRT.

From the above we conclude that looking at a single performance metric, either TNT or ANRT, is misleading. Along the same lines, VMmark identifies one system as the best performing system; for the 16-core system, this is system ranked #16, and the 48-core system, this is system ranked #1. However, a TNT versus ANRT performance characterization shows there really is a trade-off in terms of TNT versus ANRT, and one cannot firmly say that system

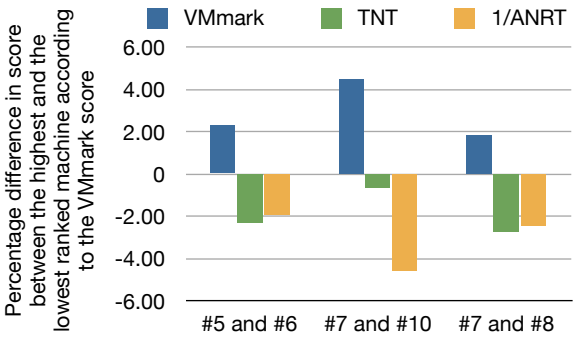


Figure 5: This graph shows three examples of ordering inversion for both TNT and 1/ANRT compared to the VMmark score. The two examples on the left compare two 32-core systems. The rightmost example compares a 32-core and a 24-core machine.

ranked #16 is better than the other systems on the Pareto frontier for the 16-core systems; the same applies for system #1 for the 48-core systems.

4.2 Scenario #2: Inversion of ordering for both TNT and ANRT

We now discuss the scenario where the VMmark score yields opposite conclusions for both the TNT and ANRT metrics. For three system comparisons, one system is ranked higher than another system according to the VMmark score, yet TNT and ANRT indicate the converse. These comparisons are illustrated in Figure 5.

Two of these comparisons are between 32-core systems: system #5 vs. system #6, and system #7 vs. system #10. Considering the VMmark score, system #5 beats system #6 by 2.3%. Considering TNT and ANRT, system #5 scores 2.4% and 2% worse compared to system #6, respectively. Similarly, system #7 yields 4.5% better consolidation performance compared to system #10 according to the VMmark score, yet it scores 4.8% and 0.6% worse in terms of ANRT and TNT, respectively.

Interestingly, the rightmost example in Figure 5 involves two systems with a different core counts: system #7 has 32 cores while system #8 has 24 cores. VMmark indicates that system #7 scores 1.8% better compared to system #8, however — and perhaps unexpectedly, given the fewer number of cores — system #8 scores 2.7% and 2.5% better than #7 in terms of TNT and ANRT, respectively.

The end conclusion from this scenario is that the VMmark score can lead to completely opposite conclusions with respect to which system yields the best consolidation performance, compared to a TNT/ANRT-based performance evaluation.

4.3 Scenario #3: Inversion of ordering for ANRT

The third scenario collects cases for which the VMmark score concurs with TNT, however, it contradicts with the ANRT score. There are many possible case studies that we can highlight that fall under this scenario. In the interest of space, we limit ourselves to comparing systems with the same number of cores — there are even more examples when comparing across different core counts.

Figure 6 shows 12 examples: 9 examples for 16-core systems, and one 24-core system example, one 32-core system example and one 48-core system example. The degree to which the VMmark score disagrees with the ANRT metric is fairly large, and the largest disagreement is observed when comparing systems ranked #31 and #37: according to the VMware score, system ranked #31 yields

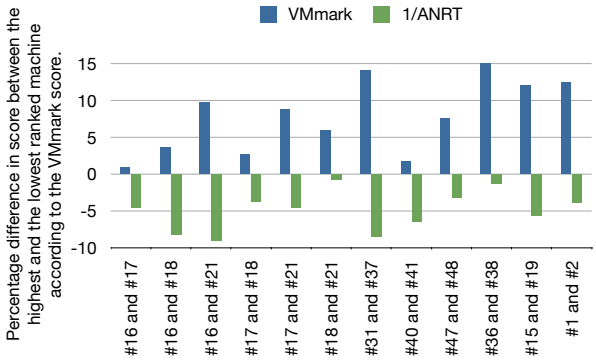


Figure 6: This graph shows examples of inversion according to the VMmark and ANRT scores. It gives the difference in score for both the VMmark and 1/ANRT metrics as a percentage of the score relative to the machine that ranks the highest according to the VMmark score. The 9 leftmost examples are for 16-core systems; the last 3 cases are for 24, 32 and 48-core systems, respectively.

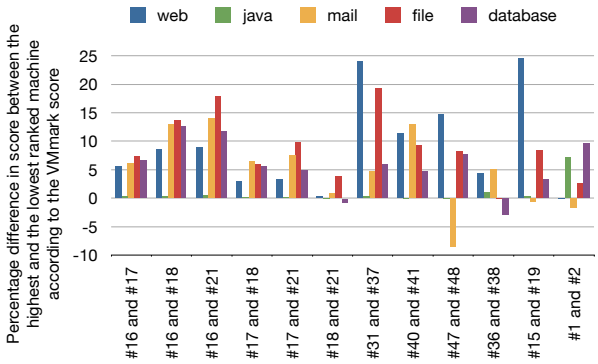


Figure 7: This graph shows the percentage in ANRT increase between the lowest ranked system and the highest ranked system for each benchmark. The 9 leftmost examples are for 16-core systems; the last 3 cases are for 24, 32 and 48-core systems, respectively.

14% better performance compared to system ranked #37, however, the increase in ANRT is as high as 9.3%.

This scenario clearly illustrates that improving system throughput through server consolidation comes at the cost of reduced per-VM performance. An interesting question now is which benchmarks suffer most from server consolidation. Figure 7 quantifies relative ANRT (on the vertical axis) for each of the benchmarks. The relative ANRT is computed as the ANRT for the system with the highest VMmark score minus the ANRT for the system with the lowest VMmark score in the case study. In other words, a positive relative ANRT means that the system with the highest VMmark score has lower per-VM performance. We find this to be the case for most of the case studies and most of the benchmarks. In addition, this graph enables understanding which benchmarks suffer most from server consolidation. Apparently, the web server, mail server, file server and database server suffer most. For some case studies, per-VM performance can be deteriorated by almost 25%. The Java server on the other hand does not seem to suffer all that much from

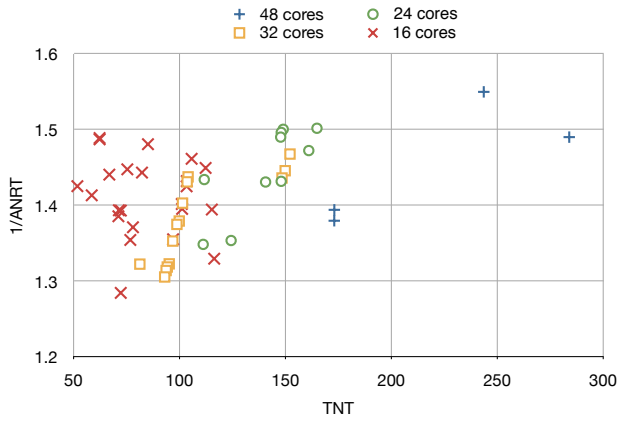


Figure 8: This graphs shows all the systems from the public VMmark data set, see Table 1, in terms of 1/ANRT versus TNT.

server consolidation. This suggests that the Java server is primarily CPU-intensive; on the other hand, the other benchmarks are more I/O-intensive and thus harder to virtualization and consolidate at low overhead.

4.4 Scenario #4: Inversion of ordering for TNT

The last scenario is supported by a single case for which the VMmark score yields an opposite conclusion compared to the TNT metric, but not for the ANRT metric. Consider the systems #41 and #42: #41 beats #42 by a margin of 0.7% in VMmark score — arguably not a very large margin. Both systems sustain 10 tiles, yet system #42 attains a drop of 8.4% in terms of TNT metric. (The difference in ANRT is almost negligible: 0.28%.) The fact that we can find only a single case in this scenario suggests that the VMmark scoring scheme leans towards weighing aggregate system throughput more heavily than per-VM performance.

4.5 Comparing all systems in terms of TNT and ANRT

Figure 8 shows all 51 systems in terms of 1/ANRT versus TNT. Here, the systems ranked #1 and #2 form the global Pareto frontier; these are the best performing systems across all the systems considered in this study. The secondary Pareto frontier consists of an interesting set of systems, namely systems ranked #3, #4, #5 and #6. Systems #3 and #4 are 48-core systems whereas #5 and #6 are 32-core systems: systems #3 and #4 achieve only slightly better TNT performance at the expense of a significant drop in per-VM performance.

4.6 ANRT versus average per-tile score

As mentioned before, VMmark reports the total aggregate tile score as its final score along with the number of tiles. The primary performance metric is the total aggregate tile score. The average per-tile score can be computed as the total aggregate tile score divided by the number of tiles. Although the average per-tile score correlates very well with ANRT (correlation coefficient of -0.99), it falls short for two reasons. First, it does not properly account for the different VMs in each tile (because of the geometric mean as argued before). Second, a TNT/ANRT characterization is more generally applicable, and allows for quantifying performance in a non-tiled setup, i.e., in a real server consolidation setup.

5. Conclusion

In spite of the current interest in virtualization and server consolidation, there is no consensus on how to benchmark consolidated servers. In this paper, we proposed two novel performance metrics, Total Normalized Throughput (TNT) and Average Normalized Reduced Throughput (ANRT), which characterize different dimensions of consolidated server performance: TNT quantifies total aggregate system performance, whereas ANRT quantifies per-VM performance. In addition, we argued that a meaningful and precise server consolidation performance characterization should use both metrics. We contrasted TNT/ANRT performance characterization against the VMmark scoring methodology using published performance numbers for a large set of commercial systems. This evaluation reveals several cases in which TNT/ANRT and VMmark come to different conclusions about which system outperforms the other. The pitfalls are that VMmark may say one system outperforms the other whereas TNT/ANRT reveals there really is a trade-off in aggregate versus per-VM performance. Or, even worse, for a few cases, VMmark yields a completely opposite conclusion with respect to which system is the best, both in terms of system throughput and per-VM performance. Finally, the VMmark score tends to weigh system throughput more heavily than per-VM performance; as a result, the VMmark scoring scheme gives a narrow view on consolidated server performance which largely omits the per-VM performance perspective. Overall, we hope this paper will help reaching consensus on how to benchmark consolidated servers, and we believe that a benchmarking methodology that scores both overall system performance and per-VM performance metrics are key in order to do a meaningful and precise performance characterization.

6. Acknowledgements

We would like to thank the anonymous reviewers for their valuable comments and suggestions. Andy Georges is supported through a postdoctoral fellowship by the Research Foundation–Flanders (FWO). Additional support is provided by the FWO projects G.0232.06, G.0255.08, and G.0179.10, and the UGent-BOF projects 01J14407 and 01Z04109.

References

[1] P. Apparao, R. Iyer, X. Zhang, D. Newell, and T. Adelmeyer. Characterization and Analysis of a Server Consolidation Benchmark. In *Proceedings of the International Conference on Virtual Execution Environments (VEE)*, pages 21–29, March 2008.

[2] J.P. Casazza, M. Greenfield, and K. Dhi. Redefining Server Performance Characterization for Virtualization Benchmarking. In *Intel Technology Journal*, Vol 10, Issue 3, pages 243–252, Aug. 2006.

[3] H. Cragon Computer Architecture and Implementation. Cambridge University Press. 2000.

[4] S. Eyerman and L. Eeckhout. System-level performance metrics for multiprogram workloads. In *IEEE Micro*, pages 42–53, May–June 2008.

[5] L. K. John. Aggregating Performance Metrics over a Benchmark Suite. In *Performance Evaluation and Benchmarking*, pages 47–58, CRC Press, 2006.

[6] K. T. Lim, P. Ranganathan, J. Chang, C. D. Patel, T. N. Mudge, and S. K. Reinhardt. Understanding and Designing New Server Architectures for Emerging Warehouse-Computing Environments. *Proceedings of the International Symposium on Computer Architecture (ISCA)*, pages 315–326, June 2008

[7] K. Luo, J. Gummaraju, and M. Franklin. Balancing Throughput and Fairness in SMT Processors. *Proceedings of the International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pages 164–171, Nov. 2001

[8] V. Makhija, B. Herndon, P. Smith, L. Roderick, E. Zamost, and J. Anderson. VMmark: A Scalable Benchmark for Virtualized Systems. *VMWare Technical Report*, VMware-TR-2006-002, Sept. 2006

[9] A. Snavely, and D. M. Tullsen. Symbiotic Jobscheduling for a Simultaneous Multithreading Processor. In *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, pages 234–244, Nov. 2000.