Performance of VMware vCenter™ 5.0 in Remote Offices and Branch Offices

Performance Study

TECHNICAL WHITE PAPER

vmware®

Table of Contents

Introduction

As VMware vCenter™ has become the most popular tool to manage VMware vSphere® ESXi™ hosts and virtual machines (VMs), customers have deployed vCenter in many networks, such as local area network (LAN) with high-bandwidth and low-latency links, and wide area network (WAN) with low-bandwidth and high-latency links. Deploying vCenter in a remote offices and branch offices (ROBO) environment is a representative use case for customers who have ESXi hosts distributed over large geographical distances. These customers choose to centralize the management of ESXi hosts and VMs by keeping only one vCenter and configuring it to manage ESXi hosts over WAN.

To help customers understand vCenter performance over WAN, we study the performance of VMware vCenter Server 5.0 in the ROBO environment in this white paper. First, to establish a baseline, we evaluated the time and bandwidth consumption of vCenter tasks over LAN (high bandwidth, low latency). Then, we evaluated the time and bandwidth consumption of vCenter tasks in five representative WAN (low bandwidth, high-latency) cases.

The experimental results show that vCenter performs well in the ROBO environment in terms of both time and bandwidth consumption. For a vCenter task, time consumption increases very slowly when network latency[1] increases by several times. Furthermore, network bandwidth is not saturated even if it decreases from 1.5 Mbps to 64 Kbps.

Executive Summary

The network traffic between vCenter and ESXi hosts can be characterized into three types: background traffic, workload traffic, and statistics traffic. The background traffic is the network communication when no vCenter task is running. The workload traffic is the network communication when vCenter tasks are executing. The statistic traffic is the network communication when vCenter retrieves statistics from hosts.

To establish a baseline, we first evaluated the time and bandwidth consumption of vCenter tasks over the LAN. Then, we evaluated the time and bandwidth consumption of vCenter tasks in low-bandwidth and high-latency cases (WAN). For each case, we first investigated the bandwidth requirement of the background traffic. Then, we studied the time and bandwidth consumption of workload traffic of representative vCenter tasks based on customer data, including host tasks, VM tasks, and vDS tasks. Finally, we measured the communication costs[2] of statistics traffic.

NETWORK PIPES	BANDWIDTH	LATENCY	PACKET ERROR RATE
Dial-up 1	64 Kbps	250 ms	0.05%
Dial-up 2	256 Kbps	250 ms	0.05%
DSL	512 Kbps	100 ms	0.05%
Satellite	1.5 Mbps	500 ms	0.10%
T1	1.5 Mbps	100 ms	0.05%
LAN	1 Gbps	1 ms	0.01%

Table 1. Network pipes – LAN and five types of WAN

---

[1] Network latency refers to delays in the propagation, transmission, or processing of the data as it travels from the origin to the destination.

[2] Communication cost refers to the size of the data sent over the network for the task to take place.

We chose five typical low-bandwidth and high-latency scenarios, which represent usage patterns of VMware customers as well as internet services provided by Internet Service Providers (ISPs). Table 1 shows the five high-latency and low-bandwidth scenarios and LAN.

Most network pipes have not only latency and limited bandwidth, but also packet corruption rate. To closely simulate these network pipes, we injected a packet error rate for each network pipe. The satellite pipe has packet error rates (PER) of between $1/10^6$ bits to $1/10^8$ bits. Considering the mean error rate of $1/10^7$ bits and an average TCP packet size of 512 bytes, the PER for the satellite pipe can be estimated to be $(512 \times 8) \div 10^7$, which approximately equals 0.05%. To err on the side of caution, we set the PER of the satellite pipe to 0.10% and the PER of other pipes to 0.05%.

The results of our tests show that vCenter is effective and efficient to manage the hosts and VMs in the ROBO environment. For a vCenter task, time consumption increases veryslowly when network latency increases by several times. For example, when network latency increases by 4 times, time consumption increases by only 1.5 times. Network bandwidth is not saturated even if it decreases from 1.5 Mbps to 64 Kbps. The bandwidth of current networks is enough for vCenter to monitor and manage hosts and VMs. The results also show that the bandwidth consumption of a host increases linearly (consistently and as expected) with the number of powered-on VMs on a host.

Evaluation Environment

Test Bed Architecture

In the experiments, we used one vCenter host with a SQL Server database to support a representative vCenter inventory in the ROBO environment. The vCenter had 640 total VMs in a configuration of 32 ESXi hosts and 20 VMs per host. The distributed resource scheduler (DRS) and high availability (HA) were enabled on the vCenter to manage these hosts and VMs. A virtual distributed switch (vDS) was created on the vCenter to manage the network configuration of hosts and VMs. Each host and its 20 VMs were connected to an exclusive port group.

Figure 1 illustrates the key components of the test bed, including the vCenter and SQL database, 32 ESXi hosts, 640 VMs, and network connections.
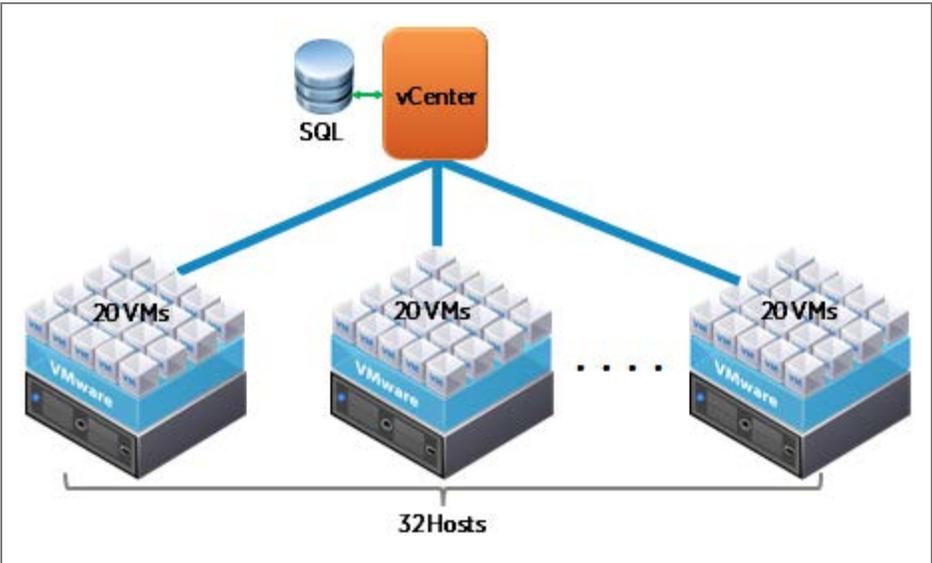

Figure 1. Experimental setup with 32 hosts and 640 VMs

To simulate the five low-bandwidth and high-latency network pipes, we limited the upload and download bandwidth for the connection between vCenter and each host, and injected network latency and packet error rate.

Hardware and Software

Identical ESXi hosts were used for this study. Each host had an eight-core 2.4GHZ Intel Xeon E5620 processor and 64GB memory.

- ESXi 5.0 GA (build 469512)
- vCenter Server 5.0 GA (build 455964)
- SQL Server 2008

Performance Metrics

We first measured the bandwidth consumption of background traffic between vCenter and each host. Second, we measured the time and bandwidth consumption of three types of vCenter tasks: host tasks, VM tasks, and vDS tasks. Third, we measured the communication costs of statistics traffic. These measurements are shown as follows:

- Background traffic
- Host tasks
    - Add and remove a host
    - Enter and exit maintenance mode
- VM tasks
    - Power on and power off a VM
    - Create and delete a snapshot
    - Create, remove, and clone a VM
    - Unregister and register a VM
    - Reset and reconfigure a VM
    - Suspend and resume a VM
- vDS tasks
    - Create and remove a vDS
    - Add and remove a static port group
- Statistics traffic

Methodology

Measuring the time consumption was straightforward: we logged the start and end times of each vCenter task.

To measure the bandwidth consumption, we first utilized TShark[3] to capture TCP packets at port 443 of the VM that vCenter resided in. For each packet, TShark logged its time stamp, size, source IP, and destination IP. Because the start and end times of each task were logged, we retrieved the packets in the TShark log file whose time stamps were between the start and end time. Note that if a task was executed on a particular host, we used the host IP address to further filter these packets. Second, we computed the sum of the sizes of these packets as the communication cost for this task. Finally, we computed the bandwidth consumption by dividing communication time from the communication cost.

---

[3] For more information about TShark, see http://www.wireshark.org/docs/man-pages/tshark.html.

Background Traffic

Background traffic is the communication between vCenter and all hosts when no vCenter task is running. Allowing background traffic to transfer between vCenter and hosts properly is the basic requirement for a vCenter to manage the hosts.

We measured the bandwidth consumption for a host with a different number of powered-on VMs. First, we powered off all VMs and used TShark to measure the communication cost for 2 minutes. Second, we consecutively powered on 20 VMs in each host and then used TShark to measure the communication cost for 2 minutes. Finally, we computed the average communication cost for 32 hosts and then calculated the average bandwidth consumption.

Figure 2 shows the average bandwidth consumption of routine communication for both network traffic directions with different numbers of powered-on VMs on each host.
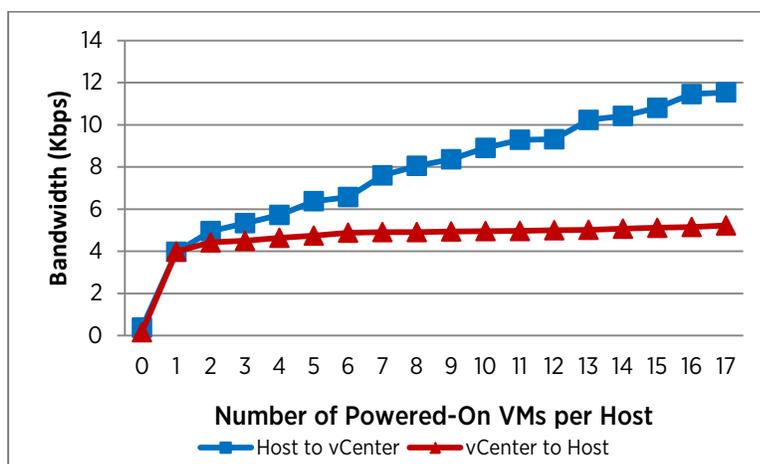
Figure 2. Bandwidth consumption of background traffic. The bandwidth increases linearly with the number of powered-on VMs in each host.

In this figure, the bandwidth consumption for both network traffic directions grow linearly with the number of powered-on VMs in each host. Therefore, we conducted linear regression on the result and concluded the following two formulas to estimate the bandwidth consumption of routing communication in our context.

$$B_{HostToVC} = 4.0 + 0.5 \times n \ (n \geq 1) \qquad (1)$$

$$B_{VCToHost} = 4.4 + 0.055 \times n \ (n \geq 1) \qquad (2)$$

In the above equations:

- $B_{HostToVC}$ is the first value you are trying to find; this is the bandwidth consumption of the host to vCenter.
- $B_{VCToHost}$ is the second value you are trying to find; this is the bandwidth consumption of vCenter to the host.
- n is the number of powered-on VMs on the host and that number must be greater than or equal to 1.

In our context, even under the lowest bandwidth network pipe, Dial-up 1, there is enough bandwidth (>14 Kbps) to support the background traffic of 20 powered-on VMs in each host.

Host Tasks

In this section, we measured the time and bandwidth consumption of four host tasks in the ROBO environment, including adding and removing a host, and entering and exiting maintenance mode. The experimental results show that vCenter efficiently handles these host tasks in the five representative low-bandwidth and high-latency networks. For a host task, the time consumption increases very slowly when network latency increases by several times. For example, when network latency increases by 4 times, the time consumption increases by only 0.8 times. The upload and download bandwidth of the five network pipes are not saturated when they decrease from 1.5 Mbps to 64 Kbps.

Add and Remove a Host

We added and removed each host and computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Note that in these tasks, hosts do not have VMs. Figure 3 shows the average time consumption and Figure 4 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.
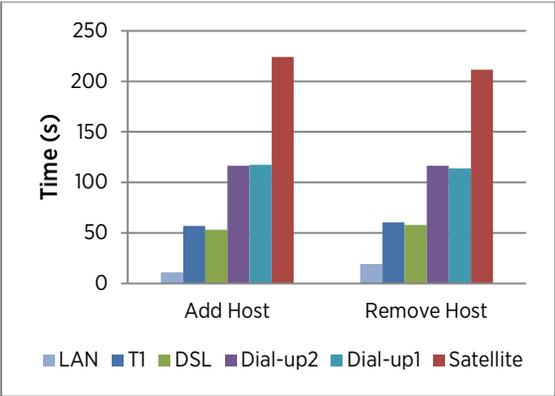

Figure 3. Time consumption for adding and removing a host. The time consumption increases slowly with network latency.

In Figure 3, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 0.85 times. The reason for this is that when vCenter executes a task, vCenter collects the process status of the test by exchanging multiple messages with hosts. The time consumption of each message exchange consists of the total network latency for both directions and the time for transferring the message.
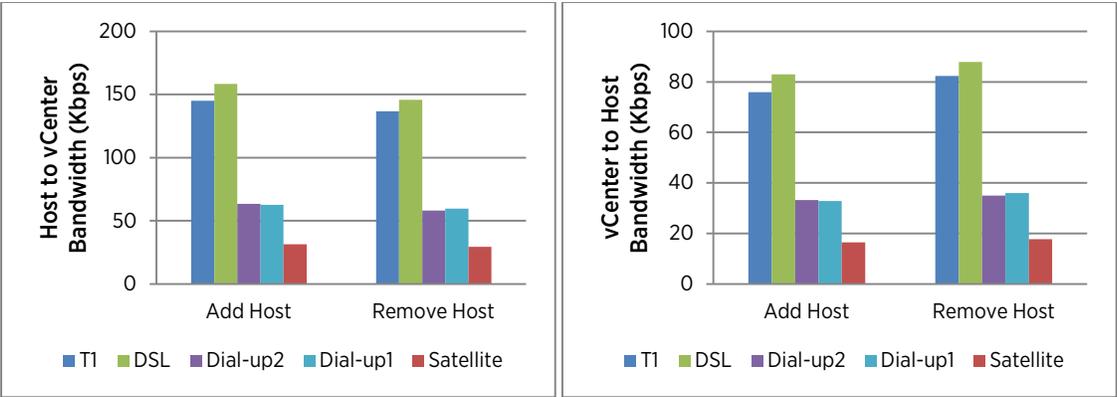

Figure 4. Bandwidth Consumption for adding and removing a host. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

Figure 4 shows that the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case (Dial-up1). Network latency is the major factor that affects the bandwidth consumption of these tasks. For the same task, Satellite network pipe always consumes the smallest bandwidth. The reason for this is that for each task, the communication costs from vCenter to a host and from a host to vCenter remain constant for different network pipes, and the time consumption increases with network latency. Due to the same reasons, the experimental results of other host tasks and VM tasks will show the similar results of time consumption and bandwidth consumption as shown in Figures 3 and 4,

Enter and Exit Maintenance

We entered and exited maintenance of each host with 20 powered-off VMs and then computed the average time consumption and bandwidth consumption. Figure 5 shows the average time consumption for these two tasks, and Figure 6 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.
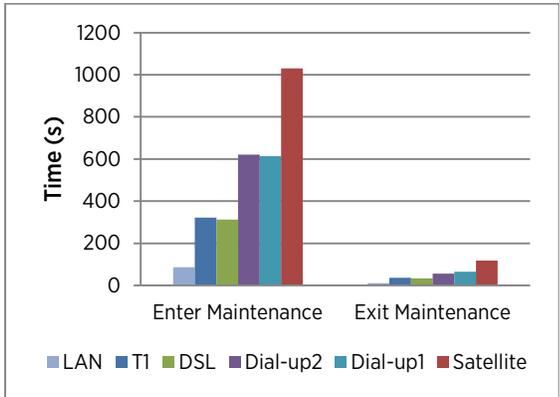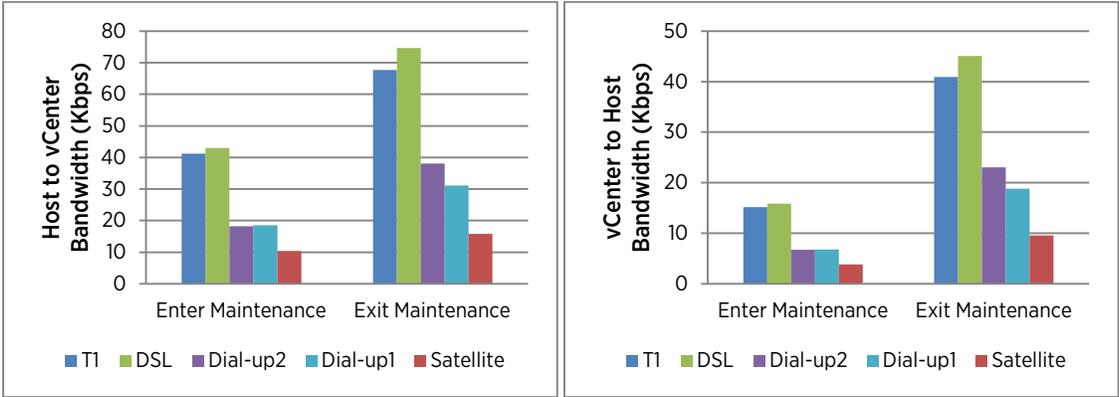

Figure 5. Time consumption for entering and exiting maintenance. The time consumption increases slowly with network latency.

In Figure 5, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 0.73 times. The reason is the same as that for adding and removing a host. When vCenter executes a task, vCenter collects the process status of the test by exchanging multiple messages with hosts. The time consumption of each message exchange consists of the total network latency for both directions and the time for transferring the message.

Figure 6. Bandwidth consumption for entering and exiting maintenance mode. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

In Figure 6, the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case, Dial-up1. For the same task, the Satellite network pipe always consumes the smallest bandwidth. The reason is that for each task, the communication costs from vCenter to a host and from a host to vCenter remain constant for different network pipes, and the time consumption increases with network latency.

VM Tasks

Next we measured the time and bandwidth consumption of thirteen VM tasks in the ROBO environment. We powered on all VMs in each host and then ran VM tasks. Each task was executed once on each host, and then we computed the average time and bandwidth consumption.

The experimental results also demonstrate that vCenter efficiently executes these VM tasks in the five low-bandwidth and high-latency networks. For a VM task, the time consumption increases much slowly when network latency increases by several times. For example, when network latency increases by 4 times, the time consumption increases by only 1.5 times. The upload and download bandwidth of the five network pipes are not saturated when they decrease from 1.5 Mbps to 64 Kbps.

Power On and Power Off a VM

We powered off and then powered on a VM on each host. Then we computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Figure 7 shows the average time consumption and Figure 8 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.
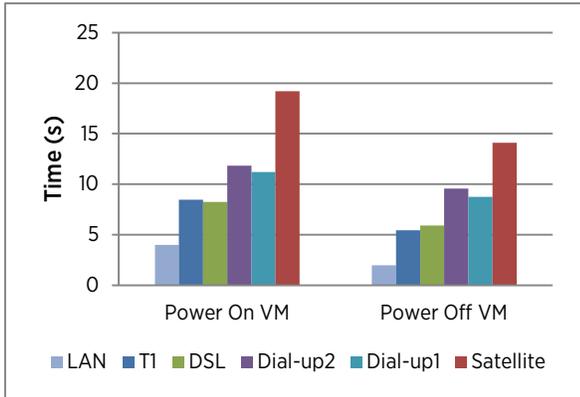

Figure 7. Time consumption for powering on and off a VM. The time consumption increases slowly with network latency.
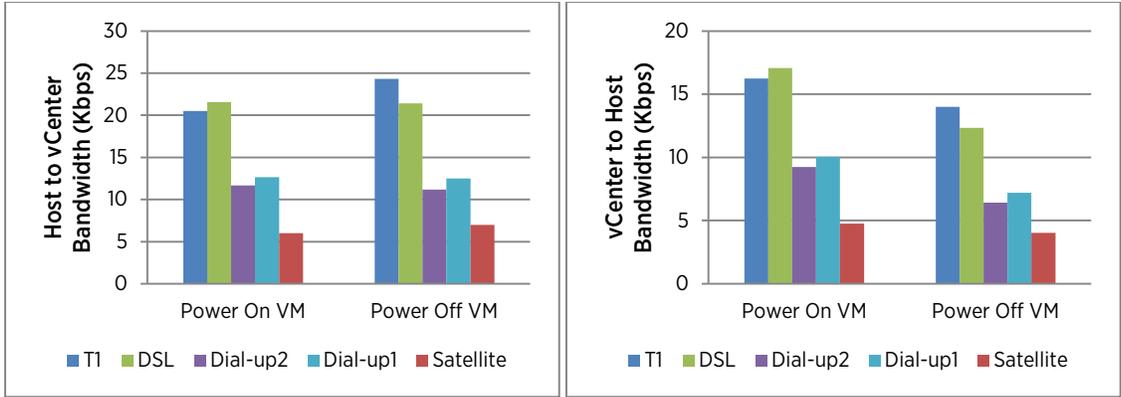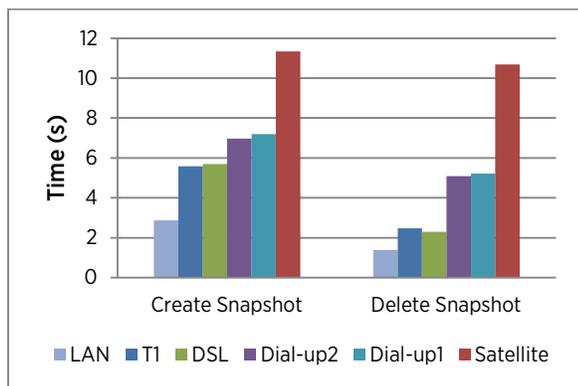

Figure 8. Bandwidth consumption for powering on and off a VM. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

In Figure 7, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 1.5 times. The network latency is the major factor that affects the time consumption of these tasks due to multiple rounds of message exchanges between vCenter and a host. In Figure 8, the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

Create and Delete a Snapshot

We created and deleted a snapshot of a VM on 32 hosts, and then computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Figure 9 shows the average time consumption and Figure 10 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.

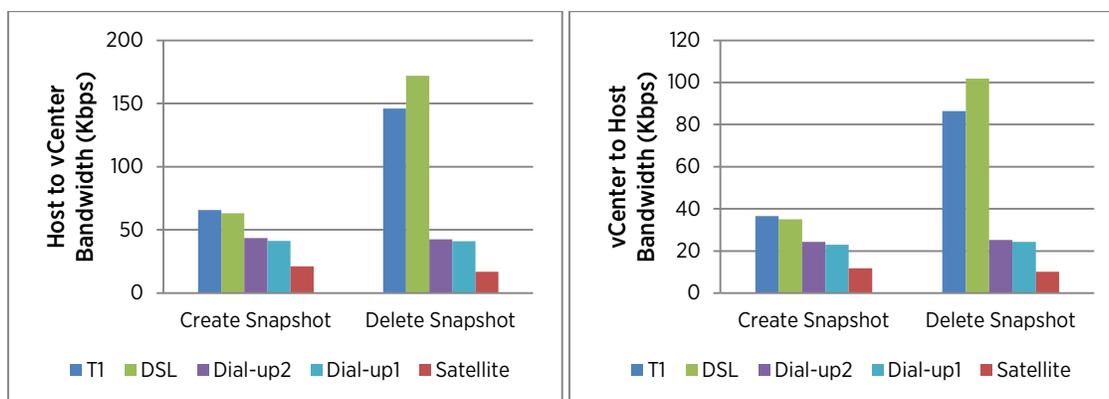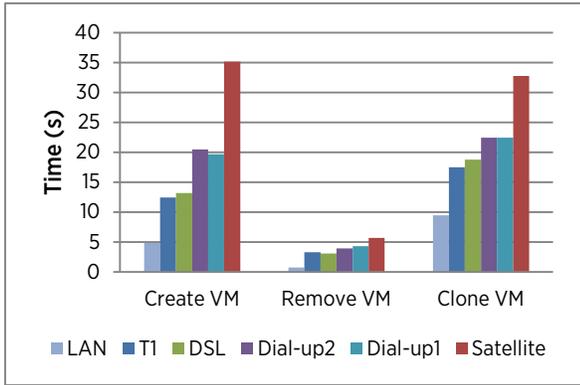**Figure 9. Time consumption for creating and deleting a snapsot. Time consumption increases slowly with network latency.**

Figure 10. Bandwidth consumption for creating and deleting a snapshot. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

These two figures show that when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 3 times, and the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

Create, Remove, and Clone a VM

We created, removed, and cloned a VM on each host, and then computed the average time and bandwidth consumption. Figure 11 shows the average time consumption and Figure 12 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.

Figure 11. Time consumption for creating, removing, and cloning a VM. The time consumption increases slowly with the network latency.
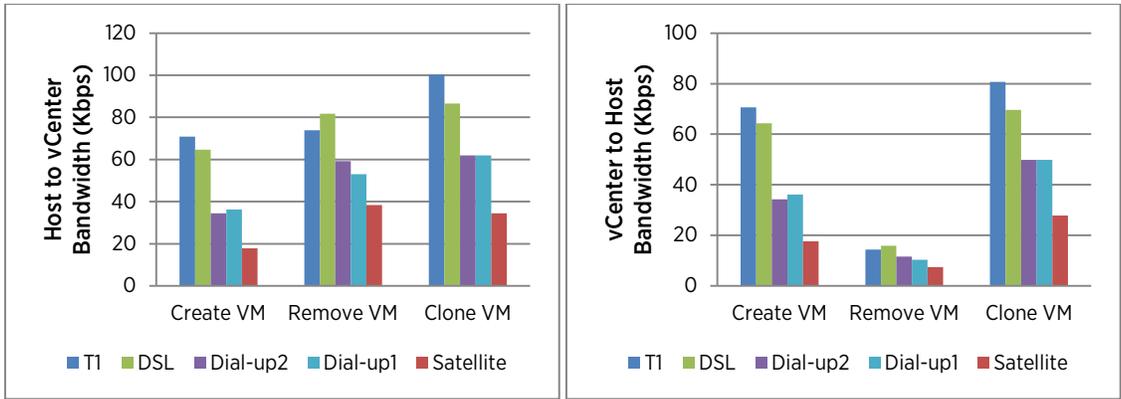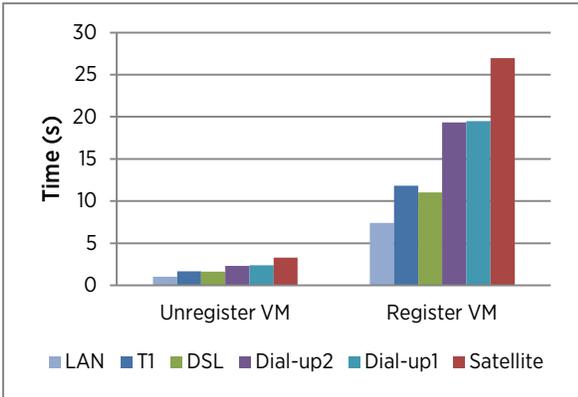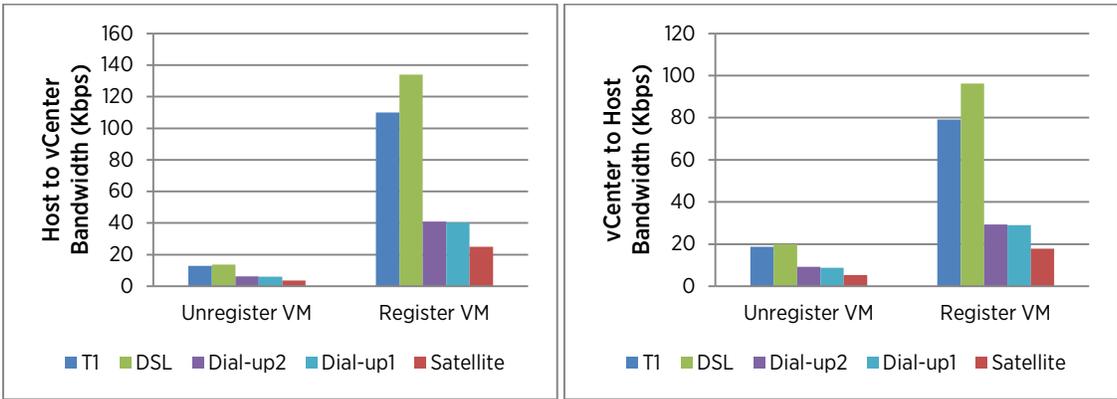

Figure 12. Bandwidth consumption for creating, removing, and cloning a VM. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

In these two figures, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 1.1 times, and the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

Unregister and Register a VM

We unregistered and registered a VM on each host, and then computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Figure 13 shows the average time consumption and Figure 14 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.
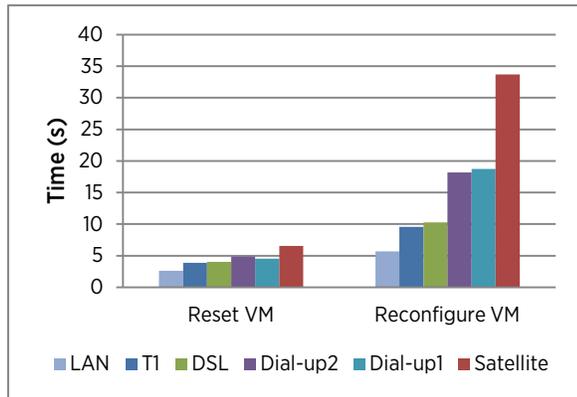

Figure 13. Time consumption for unregistering and registering a VM. The time consumption increases slowly with the network latency.

Figure 14. Bandwidth consumption for unregistering and registering a VM. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

In these two figures, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds) the time consumption of these tasks increases by only 1.1 times, and the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

Reset and Reconfigure a VM

We reset and reconfigured a VM on each host, and then computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Figure 15 shows the average time consumption and Figure 16 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.

Figure 15. Time consumption for resetting and reconfiguring a VM. The time consumption increases slowly with network latency.
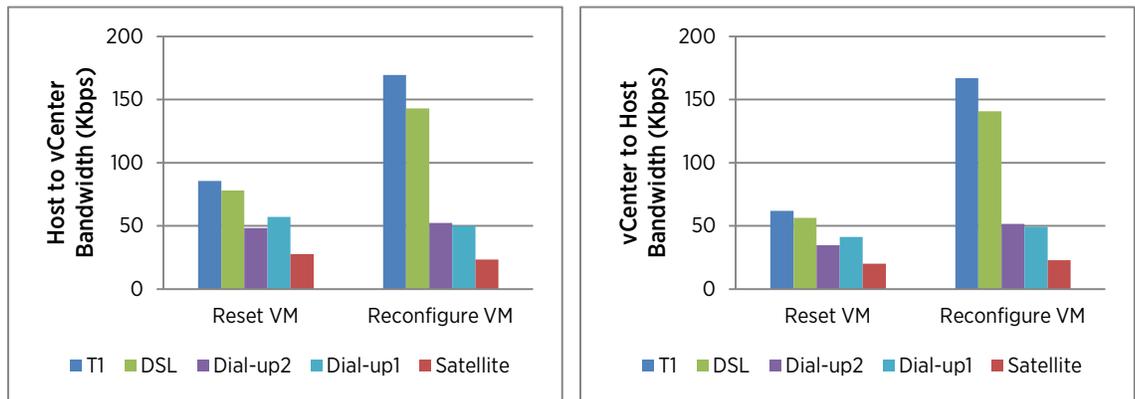

Figure 16. Bandwidth consumption for resetting and reconfiguring a VM. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

In these two figures, when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 1.6 times, and the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

Suspend and Resume a VM

We suspended and resumed a VM on each host, and then computed the average time consumption, and host-to-vCenter and vCenter-to-host bandwidth consumption. Figure 17 shows the average time consumption and Figure 18 shows the average host-to-vCenter and vCenter-to-host bandwidth consumption.
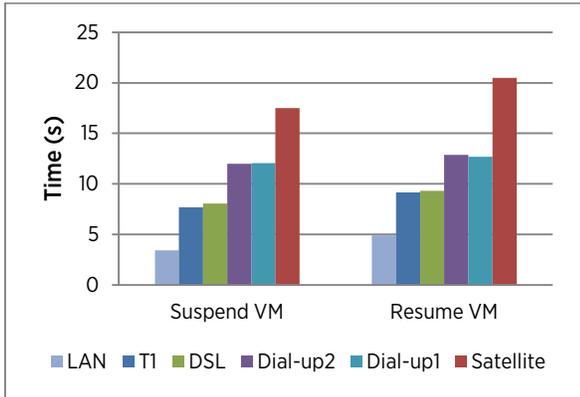

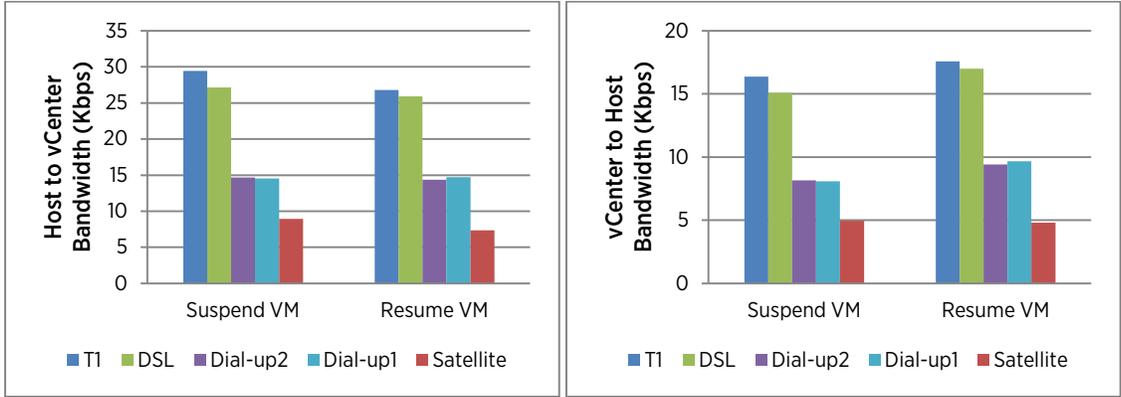**Figure 17. Time consumption for suspending and resuming a VM. Time consumption increases slowly with network latency.**

Figure 18. Bandwidth consumption for suspending and resuming a VM. The more bandwidth a network pipe has, the more bandwidth is used for a vCenter task. However, the network pipes are not saturated for all the five representative low-bandwidth cases.

These two figures show that when network latency increases by 4 times (that is, from 100 milliseconds to 500 milliseconds), the time consumption of these tasks increases by only 1.3 times, and the upload and download bandwidth of the five network pipes are not saturated even under the lowest bandwidth case.

vDS Tasks

In this section, we evaluated the performance of vDS tasks without communicating with hosts. Each vDS task was executed for 32 iterations. We measured only the average time consumption of each vDS task.
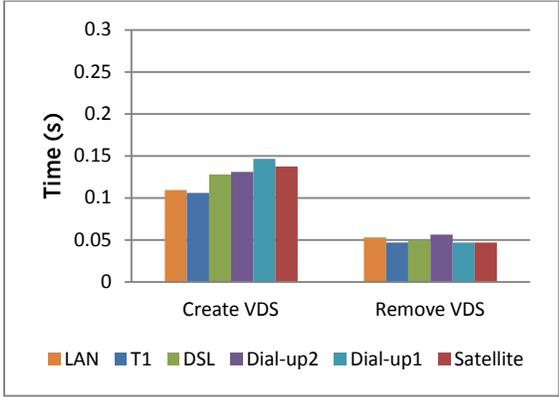

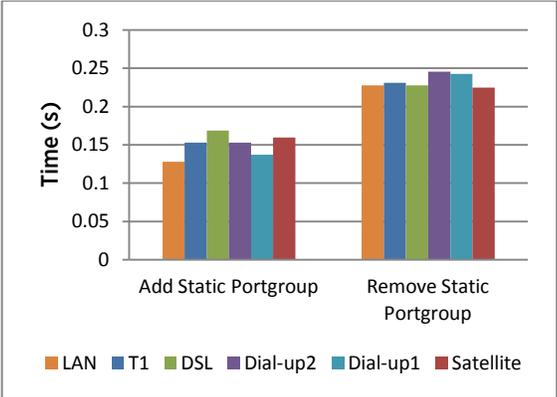**Figure 19. Time consumption for creating and removing a vDS.**

Figure 20. Time consumption for adding and removing a static port group

Figures 19-20 show that vDS tasks are significantly faster than host and VM tasks. vDS tasks are over two orders of magnitude faster than host tasks and over one order of magnitude faster than VM tasks.

Statistics

Finally, we measured the communication costs of statistics traffic. If the function of retrieving statistics information from hosts is enabled, ESXi hosts periodically send certain metrics to vCenter. One of four statistics levels, level 1 to level 4, can be enabled in vCenter. Level 1 represents the smallest set of statistics sent to the vCenter, and level 4 represents the largest set of statistics sent to the vCenter. Note that the background traffic cannot be disabled. Thus, the communication cost includes both statistics traffic and background traffic.

We evaluated the average communication costs between vCenter and a host for one hour under different statistics levels and different numbers of powered-on VMs on each host. The statistics levels are from level 1 to level 4 and the numbers of powered-on VMs on each host are 20 and 40. Figures 21-22 show the communication costs of different statistics levels for 20 and 40 VMs on each host. Figure 19 shows the comparison of overall communication costs for 20 and 40 VMs on each host.
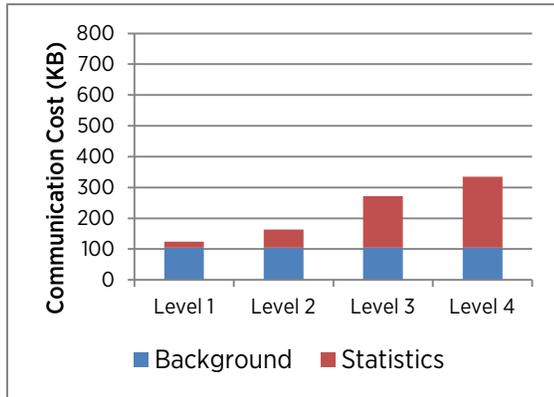
Figure 21. Communication costs of different stats levels for 20 VMs on each host. The blue bar indicates the cost of background traffic and the red bar indicates the cost of statistics traffic.
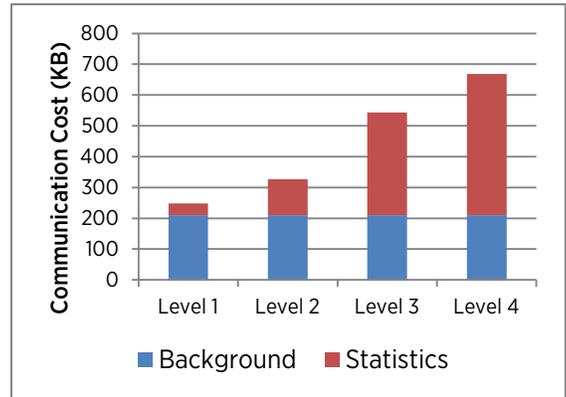
Figure 22. Communication costs of different stats levels for 40 VMs on each host. The blue bar indicates the cost of background traffic and the red bar indicates the cost of statistics traffic

These two figures show that for levels 1 and 2, the communication cost of statistics traffic are much smaller than the background traffic, which indicates that enabling level 1 or 2 statistics traffic consumes only a small bandwidth. Using statistics level 3 or 4 effectively doubles the overall traffic, even though the statistics traffic itself increases by about 3 times. The reason is that the statistics information of levels 3 and 4 include many more metrics than that of levels 1 and 2.
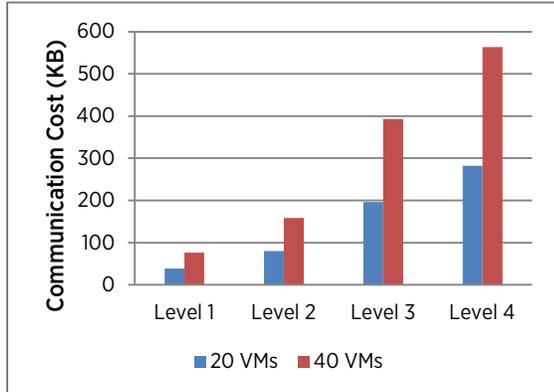

Figure 23. Comparison of overall communication costs for 20 and 40 VMs on each host

Figure 23 shows that for different numbers of VMs on each host, the size of statistics information of 40 VMs on each host is about twice larger than that of 20 VMs on each host.

Conclusion

In this paper, we investigated the performance of vCenter in the ROBO environment. First, we investigated bandwidth requirement for background traffic of vCenter. The experimental results show that the bandwidth consumption of background traffic increases linearly with the number of powered-on VMs in a host. Second, we measured time and bandwidth consumption of host tasks, VM tasks, and vDS tasks, in five low-bandwidth and high-latency network pipes. The results show that vCenter efficiently executes these tasks in terms of both time and bandwidth consumption. Third, we measured the communication costs of statistics traffic. The results show that level 1 or 2 statistics traffic consumes only a small bandwidth.

About the Author

Fei Chen is a member of technical staff in the Performance Engineering group at VMware. His work currently focuses on improving vCenter performance. He has a Ph.D. in computer science from Michigan State University.

Acknowledgements

vmware®