Microsoft RemoteFX for Remote Desktop Virtualization Host Capacity Planning Guide for Windows Server 2008 R2 Service Pack 1

Microsoft Corporation
Published: March 2011

Abstract
Microsoft RemoteFX delivers a rich user experience for session-based and virtual desktops to a broad range of client devices. This white paper is intended as a guide for capacity planning of Microsoft RemoteFX in Windows Server 2008 R2 Service Pack 1. It describes the most relevant factors that influence the capacity of a given deployment, methodologies to evaluate capacity for specific deployments, and a set of experimental results for different combinations of usage scenarios and hardware configurations.
Copyright Information

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

This White Paper is for informational purposes only. MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS DOCUMENT.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

© 2011 Microsoft Corporation. All rights reserved.

Microsoft, Active Directory, Aero, BitLocker, Excel, Hyper-V, MSDN, RemoteFX, Silverlight, Visual Studio, Windows, the Windows logo, Windows PowerShell, Windows 7, and Windows Server R2 are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries.

All other trademarks are property of their respective owners.
# Table of Contents

Introduction ........................................................................................................................................ 4

Capacity planning goals and approaches .......................................................................................... 5
  Typical evaluation approaches ........................................................................................................ 5
  Document goals ............................................................................................................................... 5
  Document non-goals ........................................................................................................................ 6
  Testing methodology ....................................................................................................................... 6
  Test setup ....................................................................................................................................... 7
  Performance testing vs. Scalability testing ...................................................................................... 8

Measurement approach ..................................................................................................................... 8

Results ............................................................................................................................................... 9
  Scalability ...................................................................................................................................... 9
  Network bandwidth usage: ............................................................................................................. 9

Adjusting resources for workload .................................................................................................... 10
  RemoteFX server system memory .................................................................................................. 10
  RemoteFX server video memory .................................................................................................... 10
  RemoteFX Processor ..................................................................................................................... 11
  RemoteFX GPU resources ........................................................................................................... 11

Appendix A: Test Hardware Details .................................................................................................. 13
  Video memory configuration: ...................................................................................................... 13
  OS and display configurations: ..................................................................................................... 13

Appendix B: Test Scenario Definitions and Workflow ....................................................................... 14
  Multimedia loads: .......................................................................................................................... 14
  Knowledge worker scenarios: ...................................................................................................... 14
  Knowledge worker + multimedia scenarios: ................................................................................ 15
**Introduction**

Microsoft® RemoteFX™ is a new feature that is included in Windows Server® 2008 R2 with Service Pack 1 (SP1). It introduces a set of end-user experience enhancements for Remote Desktop Protocol (RDP) that enable a rich desktop environment for a virtualized environment within your corporate network.

Microsoft RemoteFX enables the delivery of a full Windows® user experience to a range of client devices including rich clients, thin clients, and ultrathin clients. RemoteFX delivers a rich user experience for Virtual Desktop Infrastructure (VDI) by providing a 3D virtual adapter, intelligent codecs, and the ability to redirect USB devices in virtual machines. RemoteFX is integrated with the RDP protocol, which enables shared encryption, authentication, management, and device support. RemoteFX also delivers a rich user experience for session-based desktops and RemoteApp Programs to a broad range of client devices.

This document focuses on preliminary guidance and data around capacity planning for servers running RemoteFX for Remote Desktop Virtualization Hosts, but also briefly summarizes the facts that are equally applicable to Remote Desktop Virtualization Hosts in Remote Desktop Services deployments. For a more complete understanding of all the considerations and guidelines, it is highly recommended that you read the two architectural white papers from Microsoft on RemoteFX for RD Virtualization Host and RemoteFX for RD Session Host. The results presented in this document are based on a few scenarios that use Microsoft® Office applications. The document also provides basic guidance on the hardware and software parameters that can have a significant impact on the number of virtual machines that a server can support effectively.
Capacity planning goals and approaches

The key question capacity planning efforts try to answer is: “How many users will this server be able to host?” Other variations of the question could also be: “How much hardware is required to properly host all my users?” or “What kind of server is required to host <N> users?”

In the case of RemoteFX, the key to how many users are supported is also defined by the experience for each user. For RemoteFX, the key question implies the following additional requirement: the user experience should allow for acceptable experience in the deployment.

“Experience” and “acceptable” are relatively subjective terms and will vary based on deployment load, but in this document we will attempt to define what can be reasonably considered a good experience for a typical knowledge worker.

Typical evaluation approaches

There are several variables that determine the capacity of a specific deployment, depending on hardware resources and usage scenarios. There are practical approaches that can help reduce the estimation error to more reasonable values, and these approaches typically result in different trade-offs between effort invested and accuracy of results. For example:

1. **Setting up a test pilot.** This is a simple way of determining the capacity for a deployment. If a simple test server is set up and the usage scenario for the deployment is determined, then adjustments can be made and adjusted to meet the final requirements. Deploying a test pilot will also help analyze any issues specific to that deployment before implementing into production. However, setting up a test pilot requires dedicated hardware and software resources. In general, this approach is suitable for most first-time deployments.

2. **Create a tool set to simulate user load.** This method helps determine the load on system by building a set of simulation tools that can generate increasing load on the system. If the tools are accurate, it will determine the load level without a large error margin. An IT administrator can either create these tools themselves or use third-party tools that are written to simulate system load. We will have sample tools available that can be used by administrators after June 2011.

3. **Extrapolate the load based on the usage of a single system.** This method requires the least amount of initial effort and resources, but may not produce very accurate results. One reason for this is that it may introduce new bottlenecks due to scale: for example, network utilization will change based on the number of machines on a network, potentially causing some server delays.

Document goals

The goal of this document is to describe the performance of RemoteFX in a Remote Desktop Virtualization Host environment, in order to help plan resources required while setting up RemoteFX. This document is intended as a guide for capacity planning for a RemoteFX system running Windows Server 2008 R2 SP1 with the RD Virtualization Host role service installed.
This document contains test results for a RemoteFX deployment running on an RD Virtualization Host server.

We seek to address deployment component issues; which components are needed and how many are needed. While this is dependent on the organization implementing the deployment, as well as the kind of tasks performed, this document serves as a guideline of the requirements for some commonly-deployed tasks.

To decide on benchmarks for performance and capacity planning, it is necessary to ensure that we have a responsive system, while ensuring the hardware resources are best utilized.

We describe how an administrator can maximize system capacity and what resources can be changed to optimize deployment goals. For instance, a deployment seeking highest scalability may need to optimize resources differently from a deployment where multimedia experience is more important.

The hardware used for each test environment is detailed in Appendix A: Test Hardware Details.

**Document non-goals**


It is not a goal of this document to describe the reference architecture of a RemoteFX system. Customers should look to reference architectures from our OEM partners. Ask your existing OEM partner for their Reference Architecture, or you can visit RemoteFX: Rich End User Experience for Virtual and Session-Based Desktops ([http://www.microsoft.com/RemoteFX](http://www.microsoft.com/RemoteFX)) for a list of supported partners. All of our partners provide reference architecture details for RemoteFX deployments.

This document does not make any specific hardware recommendations, including servers, thin clients, zero clients, storage, networking or GPUs. You should contact your server OEM for recommendations on specific hardware models.

**Testing methodology**

The process of evaluating the capacity of the system is given below. There are two types of testing; system performance and system scalability:

Every section below deals with both performance and scalability. The basic process followed for both remains the same.

- Performance testing assesses factors that impact a single user experience
- Scalability testing defines how many users can be loaded in that system.

In the section Performance testing vs. Scalability testing, we elaborate how we differentiate between performance testing and scalability testing.
**Test setup**

All the tests described here were executed at Microsoft and the results were evaluated. The tests used a set of tools developed specifically for the purpose of evaluating RemoteFX capacity planning. Response times for various actions across the scenarios were used to assess the acceptable level of load under each configuration.

These tools will be made public for administrators to use on their deployments. We expect that they will be after June 2011.

The RemoteFX test configuration is shown in Figure 1. Performance and scalability tests are run in the lab on these servers. A client could be a laptop, desktop, terminal, or a thin client running RDP 7.1

![Test setup configuration](image)

*Figure 1: Test setup configuration*
**Performance testing vs. Scalability testing**

Performance testing measures the impact of each user on the system resources and applications. Performance testing follows the behavior of each program and system resource as the number of users increase. For example, we could measure the performance of a single program running in a virtual machine as the number of users of the system increase. Alternately, we could track the virtual machine’s CPU utilization as the number of virtual machines on the server increase. For performance, we track results per virtual machine, per program or per resource.

Scalability testing is done to confirm if the performance of the systems sustains over an increase in the number of users. For scalability testing, the same user scenarios per user are repeated in a loop to increase the number of users. Based on the results per user, we review the system behavior as the load on the system increases. This way, we track results on the system as a whole and measure system behavior at different loads.

**Measurement approach**

Measuring multi-media quality is important for determining the quality of codec algorithms as well as determining the overall system scalability.

Frame rate and jitter rate are the two multi-media quality metrics that are measured.

- Frame rate is the average number of frames received per second.
- Jitter rate is the number of frames that arrive with noticeable “jitter” per second.

If the time interval between arrivals of two frames is more than a specific threshold, users notice a short interruption in the media stream. This is termed “jitter”. Different users notice a different amount of jitter.

Therefore, a subjective test is needed to determine the threshold of a noticeable jitter. If the jitter a user perceives in multi-media content is low and video/audio continues to remain in sync, the overall quality may be acceptable. Therefore, for our tests, we measure and track the number of frames that arrived with noticeable jitter and define a quality bar based on it.

Figure 2 illustrates how jitter rate is calculated. Let’s assume there is a multi-media scenario where 60 frames are supposed to be presented in two seconds. Among these frames, one arrives with noticeable jitter. In that case, the jitter rate is 1/2 (or 0.5).

![Figure 2: Calculating Jitter Rate](image)

It is important to determine a jitter rate that is still considered acceptable by users.
Optimization for effective frame rate and jitter rate:

When connected to the virtual machine using RDP, some of the frames arrive too close to each other and are presented as one frame to the user. In a client OS where Desktop Window Manager (DWM) is enabled, the rate of presentation is 60 fps (or every 15ms). Therefore, in our approach, we optimize by calculating frame rate and jitter rate based on the assumption that all frames arriving in 15ms intervals are presented as one frame.

Results

Scalability

The conclusion of the scalability tests was that 2 GPUs, with a video memory of 2048 MB each, were capable of scaling to up to 32 virtual machines. A single GPU, supports up to 16 virtual machines. Our testing shows a linear scale as GPUs are added.

Our scalability tests were run on the ATI v7800p GPU card. These tests include three scenarios: knowledge worker, knowledge worker with multimedia and multimedia-only (see scenario descriptions in Appendix B: Test Scenario Definitions and Workflow). All tests were run with a resolution set to 1280 X 1024.

<table>
<thead>
<tr>
<th>Scenario name</th>
<th>Resolution (1280 X 1024)</th>
<th>Number of users supported with 1 GPU (2048 MB)</th>
<th>Number of users supported with 2 GPUs (4096 MB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge worker only</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Knowledge worker with</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>Multimedia</td>
<td>16</td>
<td>16</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 1: Users Supported Per Scenario

Scalability results based on ATI GPU cards used:

<table>
<thead>
<tr>
<th>No. of GPUs</th>
<th>ATI v7800p – users supported</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 GPU</td>
<td>16</td>
</tr>
<tr>
<td>2 GPU</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 2: Users Supported Per GPUs

This is a sample representation of our results, based on the configurations above. Results will vary with different configurations.

Network bandwidth usage:

The network bandwidth recommendation is 10Mbps per user (i.e. per VM).

In our tests running the knowledge worker + multimedia scenario, we found the network bandwidth usage to be far lower, at 2-3 Mbps per user.

Network bandwidth usage ranges from 0.5Mb/s to 9.5Mb/s, and depends on the application. In tests running the knowledge worker scenario, bandwidth consumption per user is lower.
**Adjusting resources for workload**

It is important to review the resources used by the server and identify areas that could be optimized. This section talks briefly about managing server resources in order to improve performance.

Based on usage scenarios and hardware configuration, the deployment capacity will vary. For example, some points to consider while planning your deployment are:

- Number of GPUs on your system
- Video memory capacity on the graphics cards
- Processor and hardware resources of your system, etc.

The system resources are listed below, with possible optimizations where applicable.

**RemoteFX server system memory**

For every virtual desktop, RemoteFX uses system memory both, in the guest operating system as well as the RemoteFX server itself. The hypervisor guarantees the availability of system memory for a guest operating system. On the RD Virtualization Host server each RemoteFX-enabled virtual desktop needs to advertise its system memory requirement to the hypervisor. When the RemoteFX-enabled virtual desktop is starting, the hypervisor reserves additional system memory on the RemoteFX server for a RemoteFX-enabled virtual desktop.

The RemoteFX memory requirement is dynamic because the amount of memory consumed on the RemoteFX server is dependent on the number of monitors associated with the RemoteFX-enabled virtual desktop and the maximum resolution for those monitors.

**RemoteFX server video memory**

Every RemoteFX-enabled virtual desktop uses the host video memory on the GPU to render the desktop. In addition to rendering, the video memory is also used by the codec to compress the rendered screen. The amount of memory needed is directly based on the amount of monitors provisioned to the virtual machine.

The video memory allocated varies based on the number of monitors and the system screen resolution. System experience differs based on screen resolution. Some users may require a higher screen resolution for specific tasks. There is greater scalability with lower resolution settings, with all other settings constant as mentioned in the hardware configuration table. Table 3 describes this.
Virtual machine resolution | Virtual Processors/RAM | Number of GPUs | Number of virtual desktops *
--- | --- | --- | ---
1024 X 768 | 2 GB | 1 GPU | 27
1280 X 1024 | 2 GB | 1 GPU | 16
1600 X 1200 | 2 GB | 1 GPU | 11

Table 3: Example of Scalability Versus Video Resolution

* These are theoretical numbers, indicating how many RemoteFX-enabled virtual desktops fit on a GPU with a given amount of video memory. They do not take into account the amount of GPU resources consumed by each virtual machine. The GPU resource consumption will vary from workload to workload. Actual numbers will vary based on the workloads executing in the virtual machines.

**RemoteFX Processor**

In the case of CPU resources, the hypervisor schedules the RemoteFX server as well as the RemoteFX-enabled virtual desktops. Unlike system memory, there isn’t additional resource related information that RemoteFX needs to share with the hypervisor. The additional overhead that RemoteFX brings into the RemoteFX-enabled virtual desktop is the execution of the virtual GPU driver and a user-mode RDP stack.

On the RemoteFX server side, the overhead is increased, as the system runs an additional process (rdvgm.exe) per RemoteFX-enabled virtual desktop. This process uses the graphics device driver to execute commands on the GPU. The codec also uses the CPUs for compressing the screen data that needs to be sent back to the client.

More virtual processors mean better user experience. This is particularly pronounced with the video experience.

**RemoteFX GPU resources**

For every RemoteFX-enabled virtual desktop there is a corresponding DirectX process running on the RemoteFX server. This process replays all the graphics commands it receives from the RemoteFX-enabled virtual desktop on to the physical GPU. For the physical GPU it is equivalent to simultaneously running multiple DirectX applications. Typically, graphics devices and drivers are tuned to execute a few applications on the desktop. RemoteFX stretches the GPUs to be used in a unique manner. Since RemoteFX is a new type of workload on graphics devices, the parameters that need to be measured are unique to it. RemoteFX does the following operations on the GPU:
1. It renders the desktop on the RemoteFX-enabled virtual desktop.
2. It runs the first stage of the compression algorithm on the rendered content.
3. It reads back the partially compressed content from the GPU into the system and then does the final stage of compression on the CPU.

In order to measure how the GPU is performing on a RemoteFX server, performance counters have been added to measure the GPU response to RemoteFX requests. Usually when a GPU resource is low on resources, read and write operations to the GPU take a long time to complete. By using performance counters, administrators will be able to take preventive action, eliminating the possibility of any downtime for their end users.

**Optimization for scalability:** Every additional GPU card improves system scalability (for e.g. by about 16 users with every ATI v7800p card, as illustrated in table 3). Depending on the type of server, the administrator can increase the number of GPU cards. At the time of the writing of this document, up to 2 GPU cards can be used with some servers.
Appendix A: Test Hardware Details

Video memory configuration:
Definition: GPU stands for Graphics Processing Unit – other names used are video controllers and graphics adapters.

There were two different GPU configurations tested, one with 4GB video memory and another with 2GB video memory on the graphics card.

The test setup used the ATI v7800p GPU card. For an updated list of supported GPU configurations, see the MSDN posting: http://blogs.msdn.com/b/rds/archive/2010/07/08/more-partner-momentum-around-microsoft-remotefx-in-windows-server-2008-r2-sp1-beta.aspx


The full list of hardware requirements for RemoteFX can be found here.

OS and display configurations:
The supported operating system on servers is Microsoft Windows Server 2008 R2 SP1 (64-bit).

The supported operating systems on virtual machines are Windows 7 Enterprise SP1 and Windows 7 Ultimate SP1 (both 32-bit and 64-bit). On the client side, the remote client should be running RDP 7.1. This includes Windows 7 Service Pack 1 for rich clients, Windows Embedded Standard 7 and Windows Embedded Standard 2009 for thin clients as well as partners that offer RDP 7.1 on their own OS platform.
Appendix B: Test Scenario Definitions and Workflow

There are three different loads used in the test process. While the scalability of the system is not altered based on the loads, these are given here because they do cause variations in the system such as reliability and user experience. In the section “Adjusting resources for workload” we describe how administrators can maximize the deployment performance based on the criteria of their choice.

Multimedia loads:
The graphics scenarios consist of workloads that include Silverlight and Flash applications, in addition to MS Office applications.

Sample applications run to test graphics scenarios could include Microsoft PowerPoint, Windows Media player, Silverlight and Flash.

<table>
<thead>
<tr>
<th>Microsoft Graphics Technologies</th>
<th>Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDI</td>
<td>Microsoft Office 2003/2007</td>
</tr>
<tr>
<td></td>
<td>Internet Explorer</td>
</tr>
<tr>
<td></td>
<td>Adobe Flash</td>
</tr>
<tr>
<td>D3D 9 and below</td>
<td>Windows Media Player</td>
</tr>
<tr>
<td></td>
<td>Microsoft Silverlight</td>
</tr>
</tbody>
</table>

Table 4: Applications Run on Graphics

In our tests, a multimedia-only load includes Windows Media player, Microsoft PowerPoint (with animation) and Flash, one at a time, run again an increasing number of users.

Knowledge worker scenarios:
Typing Speed = 35 words per minute

Definition: The Knowledge Worker scenario includes typical business tasks involving Microsoft Office and Internet Explorer. It includes creating and saving Word documents, printing Excel spreadsheets, communicating by e-mail in Outlook, adding slides to PowerPoint presentations, running slide shows, and browsing Web pages in Internet Explorer.

The load on the server is defined by the deployed applications, the hardware configuration, the system software configuration, and the user interaction, which can differ substantially from deployment to deployment. While one deployment may host a relatively lightweight application that users access infrequently with low resource costs (like a data entry application), another may host a very demanding graphics-intensive application requiring a lot of resources. Below, we review different approaches that are used in evaluating the load.

Table 5 outlines sample actions that are used to test knowledge worker scenarios.
Table 5: Sample Actions use in Worker Scenarios

For scalability tests, an application scenario will be composed of several simple scenarios presented in above. The scenario is executed in a loop for each user. A proposed scenario is as follows:

- Typing a document
- Opening and scrolling a word document
- A PowerPoint slide show
- Playing a Flash clip inside IE
- Playing a Silver Light video inside IE
- Playing a short video with Windows Media Player

Knowledge worker + multimedia scenarios:
This is a combination of the knowledge worker scenario and a flash-based application, run against an increasing number of users.