

Infrastructure Planning   
and Design

Microsoft Exchange Server 2010  
with Service Pack 1

Version 1.1

Published: December 2010

Updated: July 2011

For the latest information, please see [www.microsoft.com/ipd](http://www.microsoft.com/ipd)

Copyright © 2011 Microsoft Corporation. All rights reserved. Complying with the applicable copyright laws is your responsibility.  By using or providing feedback on this documentation, you agree to the license agreement below.

If you are using this documentation solely for non-commercial purposes internally within YOUR company or organization, then this documentation is licensed to you under the Creative Commons Attribution-NonCommercial License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc/2.5/> or send a letter to Creative Commons, 543 Howard Street, 5th Floor, San Francisco, California, 94105, USA.

This documentation is provided to you for informational purposes only, and is provided to you entirely "AS IS".  Your use of the documentation cannot be understood as substituting for customized service and information that might be developed by Microsoft Corporation for a particular user based upon that user’s particular environment. To the extent permitted by law, MICROSOFT MAKES NO WARRANTY OF ANY KIND, DISCLAIMS ALL EXPRESS, IMPLIED AND STATUTORY WARRANTIES, AND ASSUMES NO LIABILITY TO YOU FOR ANY DAMAGES OF ANY TYPE IN CONNECTION WITH THESE MATERIALS OR ANY INTELLECTUAL PROPERTY IN THEM.

Microsoft may have patents, patent applications, trademarks, or other intellectual property rights covering subject matter within this documentation.  Except as provided in a separate agreement from Microsoft, your use of this document does not give you any license to these patents, trademarks or other intellectual property.

Information in this document, including URL and other Internet Web site references, is subject to change without notice. Unless otherwise noted, the example companies, organizations, products, domain names, e-mail addresses, logos, people, places and events depicted herein are fictitious.

Microsoft, Active Directory, ActiveSync, Forefront, Lync, Outlook, and Windows Server are either registered trademarks or trademarks of Microsoft Corporation in the United States and/or other countries and regions.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

You have no obligation to give Microsoft any suggestions, comments or other feedback ("Feedback") relating to the documentation. However, if you do provide any Feedback to Microsoft then you provide to Microsoft, without charge, the right to use, share and commercialize your Feedback in any way and for any purpose.  You also give to third parties, without charge, any patent rights needed for their products, technologies and services to use or interface with any specific parts of a Microsoft software or service that includes the Feedback.  You will not give Feedback that is subject to a license that requires Microsoft to license its software or documentation to third parties because we include your Feedback in them.

Contents

[The Planning and Design Series Approach 1](#_Toc297805463)

[Introduction to the Microsoft Exchange Server 2010 with SP1 Guide 2](#_Toc297805464)

[Step 1: Define the Business and Technical Requirements 7](#_Toc297805465)

[Step 2: Define the Instances of Exchange Server 2010 10](#_Toc297805466)

[Step 3: Design the Mailbox Server Infrastructure 12](#_Toc297805467)

[Step 4: Design the Client Access Server Infrastructure 20](#_Toc297805468)

[Step 5: Design the Hub Transport Server Infrastructure 27](#_Toc297805469)

[Step 6: Design the Edge Transport Server Infrastructure 34](#_Toc297805470)

[Step 7: Design the Unified Messaging Server Infrastructure 40](#_Toc297805471)

[Step 8: Define the Active Directory Domain Services Requirements 45](#_Toc297805472)

[Conclusion 46](#_Toc297805473)

[Appendix A: Job Aids 47](#_Toc297805474)

[Appendix B: IPD in Microsoft Operations Framework 4.0 53](#_Toc297805475)

[Appendix C: Microsoft Exchange Server 2010 in Microsoft Infrastructure Optimization 54](#_Toc297805476)

[Version History 55](#_Toc297805477)

[Acknowledgments 56](#_Toc297805478)

# 

# The Planning and Design Series Approach

This guide is one in a series of planning and design guides that clarify and streamline the planning and design process for Microsoft® infrastructure technologies.

Each guide in the series addresses a unique infrastructure technology or scenario. These guides include the following topics:

* Defining the technical decision flow (flow chart) through the planning process.
* Describing the decisions to be made and the commonly available options to consider in making the decisions.
* Relating the decisions and options to the business in terms of cost, complexity, and other characteristics.
* Framing the decision in terms of additional questions to the business to ensure a comprehensive understanding of the appropriate business landscape.

The guides in this series are intended to complement and augment the product documentation. It is assumed that the reader has a basic understanding of the technologies discussed in these guides. It is the intent of these guides to define business requirements, then align those business requirements to product capabilities, and design the appropriate infrastructure.

## Benefits of Using This Guide

Using this guide will help an organization to plan the best architecture for the business and to deliver the most cost-effective enterprise messaging technology.

Benefits for Business Stakeholders/Decision Makers:

* Most cost-effective design solution for an implementation. Infrastructure Planning and Design (IPD) eliminates over-architecting and overspending by precisely matching the technology solution to the business needs.
* Alignment between the business and IT from the beginning of the design process to the end.

Benefits for Infrastructure Stakeholders/Decision Makers:

* Authoritative guidance. Microsoft is the best source for guidance about the design of Microsoft products.
* Business validation questions to ensure the solution meets the requirements of both business and infrastructure stakeholders.
* High-integrity design criteria that includes product limitations.
* Fault-tolerant infrastructure, where necessary.
* Proportionate system and network availability to meet business requirements. Infrastructure that is sized appropriately to meet business requirements.

**Benefits for** Consultants or Partners:

* Rapid readiness for consulting engagements.
* Planning and design template to standardize design and peer reviews.
* A “leave-behind” for pre- and post-sales visits to customer sites.
* General classroom instruction/preparation.

Benefits for the Entire Organization:

Using this guide should result in a design that will be sized, configured, and appropriately placed to deliver a solution for achieving stated business requirements, while considering the performance, capacity, manageability, and fault tolerance of the system.

# Introduction to the Microsoft Exchange Server 2010 with SP1 Guide

Microsoft Exchange Server 2010 is an enterprise-class messaging system that provides a complete end-to-end solution for electronic messaging both within and outside of an organization. It provides a platform for users to exchange traditional email messages as well as non-traditional messages such as faxes and voice mail in a fault-tolerant, distributed capable environment that can communicate with other messaging systems across the world.

As with all enterprise-class solution deployments, Exchange Server 2010 requires proper planning around key critical areas: placement of individual servers and roles, capacity planning, performance planning, fault tolerance, and hardware configuration. To develop and implement a successful Exchange Server 2010 architecture, decisions must be made relative to each of these areas to weigh the value of implementing supporting services against the cost of managing and maintaining those services.

The purpose of this guide is to help messaging architects understand the critical architectural decisions within a design for Exchange Server 2010 with Service Pack 1 (SP1).

## Assumptions

To limit the amount of material covered in this guide to an appropriate level, the following assumptions have been made:

* The decision to implement Exchange Server 2010 has already been made.
* A functional Active Directory® Domain Services (AD DS) infrastructure that meets the requirements of Exchange Server 2010 is either in place, or planning for this has been completed. Refer to *Planning Active Directory* at <http://technet.microsoft.com/en-us/library/bb123715.aspx>. The design of AD DS is covered in the *Infrastructure Planning and Design Guide for Active Directory Domain Services* at <http://go.microsoft.com/fwlink/?LinkId=157704>.
* Because it is impractical in one document to describe every type of deployment scenario, this guide is based around a new implementation of Exchange Server 2010 as the only messaging service in the organization. Concepts provided are intended to educate the reader on the proper design and planning of an Exchange Server 2010 messaging architecture and not to provide prescriptive guidance on individual scenarios (such as migrations). This guide can help the reader validate a desired end-state messaging architecture by providing a clearly defined path that ensures a properly functioning Exchange Server 2010 infrastructure.
* The reader has knowledge in the email and electronic messaging disciplines in general, as well as experience with Windows Server®, Active Directory Domain Services, and associated technologies.
* If implementing Unified Messaging services, the reader is responsible for all knowledge beyond the implementation of the Unified Messaging server role, such as Private Branch eXchange (PBX) gateways.

## Exchange Server 2010 Design Process

This guide describes the process of planning an Exchange Server 2010 messaging infrastructure. It addresses the following items:

* Defining the goals and scope of the project by identifying appropriate features and functionality required by the business.
* Mapping the required features to individual Exchange Server 2010 server roles.
* Designing an Exchange Server 2010 messaging infrastructure that provides internal and/or external communications capabilities.
* Determining all high-availability requirements and designing and placing their applicable components.

The guide addresses the following decisions and/or activities that need to occur in planning an Exchange Server 2010 infrastructure. The eight steps below represent the most critical design elements in a well-planned Exchange Server 2010 design:

* Step 1: Define the Business and Technical Requirements
* Step 2: Define the Instances of Exchange Server 2010
* Step 3: Design the Mailbox Server Infrastructure
* Step 4: Design the Client Access Server Infrastructure
* Step 5: Design the Hub Transport Server Infrastructure
* Step 6: Design the Edge Transport Server Infrastructure
* Step 7: Design the Unified Messaging Server Infrastructure
* Step 8: Define the Active Directory Domain Services Requirements

Some of these items represent decisions that must be made. Where this is the case, the guide provides a corresponding list of common response options.

Other items in this list represent tasks that must be carried out. These types of items are addressed because their presence has major impact on the infrastructure design.

Figure 1 provides a graphical overview of the steps in designing an Exchange Server 2010 messaging infrastructure.

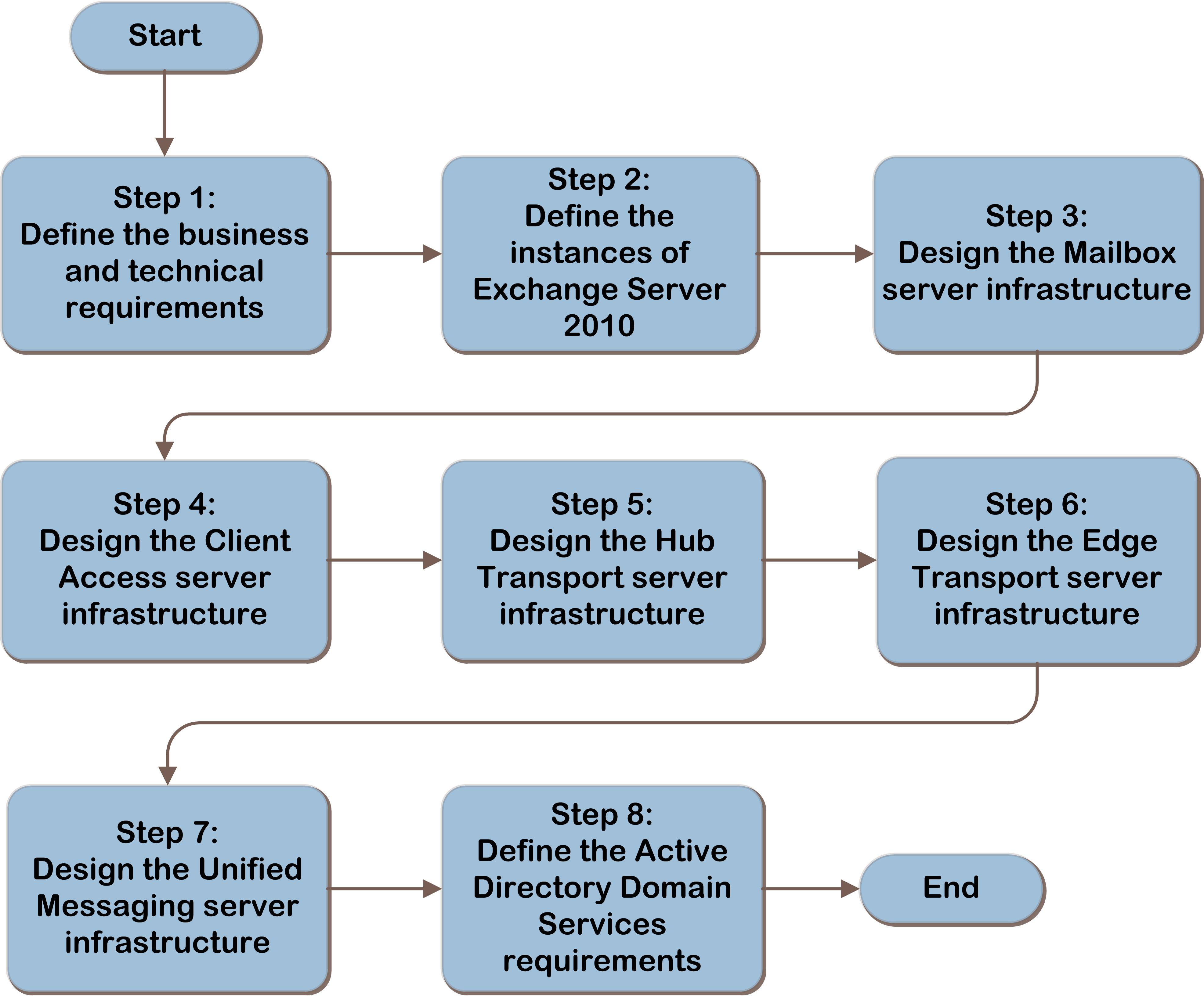


Figure 1. The Microsoft Exchange Server 2010 infrastructure decision flow

An example of an Exchange Server 2010 architecture is shown in Figure 2.

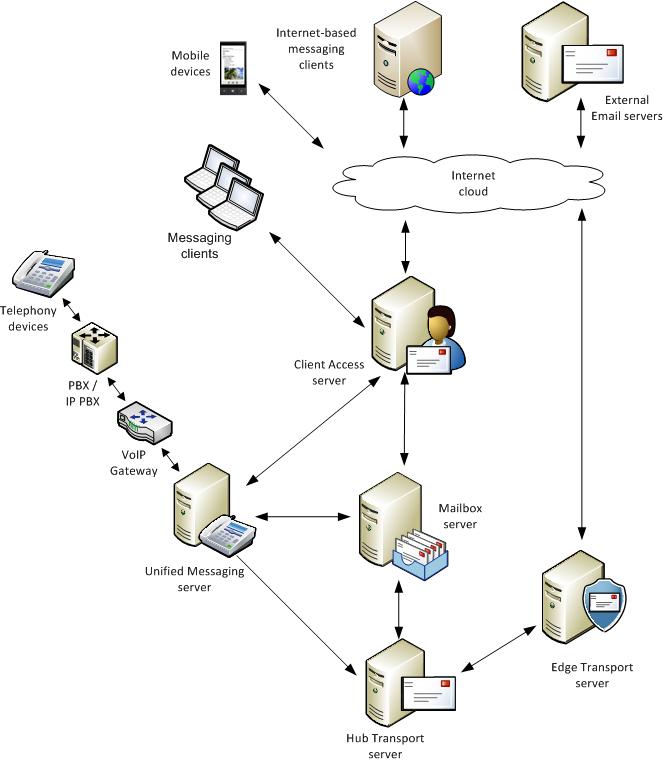


Figure 2. Example Exchange Server 2010 architecture

## Applicable Scenarios

This guide addresses considerations relative to planning and designing the necessary components for a successful Exchange Server 2010 messaging infrastructure:

* Organizations implementing a new messaging infrastructure for both internal and external communications.
* Organizations wanting to validate a new messaging infrastructure prior to planning an upgrade or migration.

## Out of Scope

The following items are considered out of scope for this guide:

* Designing coexisting messaging services (for example, Exchange Server 2003 and Exchange Server 2010 in the same organization).
* Business and/or technical decisions involved in choosing a messaging platform.
* Migrations, upgrades, or other non-green field implementations.
* External hardware configurations (for example, PBX, IP PBX, and so on) that may be required for extended functionality and are not part of a Microsoft product.
* Designing a Microsoft Lync™ 2010 architecture.

# 

# Step 1: Define the Business and Technical Requirements

In this step, the business and technical requirements will be identified. This information will be used to define the scope of the project; however, it will also be used in subsequent steps in the project to design the infrastructure.

The tasks to be performed in this step are:

1. Identify business requirements.
2. Identify technical requirements.

The result of this step will be a well-defined project scope that aligns with the business and technical requirements and sets the tone and depth for the entire project.

## Task 1: Identify Business Requirements

In this task, questions are asked of the business decision makers in the organization to determine the scope and requirements of the messaging infrastructure. The results are used to drive technical requirements and constraints on the messaging system.

### Capabilities and Scope Requirements

The following questions should be asked of the business:

* Which of the following capabilities are needed? The choices made here help determine which roles and features need to be implemented.
* **Internal email services.** Will the messaging system be used for communications with users inside the organization?
* **External email services.** Will the messaging system be used for communications with users outside of the organization?
* **Connectivity for external users.** Will the organization provide the ability for users to connect to the messaging system from outside of the organization’s firewall?
* **Public folders.** Will public folders be required for users in the organization?
* **Unified Communications (voice mail) services.** Will the messaging system be used to provide voice messaging to users in the organization?
* What portion of the organization is in scope? For example, if the company is a multinational entity with multiple subsidiaries, are all of these in scope or only part of them? Does the project include the entire company or only a subset?
* **For the parts of the organization that are in scope, do any business isolation requirements exist?** For example, do any business areas require complete separation from the rest of the organization, in effect operating as individual entities? Items to discuss include:
* **Legal isolation.** Do specific legal requirements necessitate that the organization remains isolated?
* **Political environment.** Do specific political implications require the organization to be separated?
* **Regulatory requirements.** Do specific regulatory requirements prevent the organization from being considered a single entity?
* **Business policy.** Do internal business policies require the organization to remain separated?
* **Does the organization require an operational consolidation?** For example, if the current organization is the result of multiple mergers and/or acquisitions, is there a requirement to consolidate redundant organization-wide services such as messaging?
* **Does the organization have a requirement for a unified global address list (GAL)?** For example, is there a requirement that all employees are visible in a single address list by everyone in the organization? This requirement, when coupled with multiple instances of the messaging platform, could drive the need for the implementation of additional technology.
* **Would any business policies impact messaging usage**? Document any impacts that policies might have on the following:
* **Message size limits.** How large are messages that users are allowed to send?
* **Mailbox size limits.** How large are mailboxes allowed to be?
* **Employee differentiation.** Is there a requirement that specific groups of users in the organization (for example, CxO-level executives and information workers) have different messaging limits?
* **Does the organization require internal message hygiene services?** Will the organization perform its own *message hygiene*—removal of viruses and unsolicited commercial email from the mail flow before it reaches the end user—within the organization, or will this be outsourced to a hosted message hygiene provider, or a combination of both? Legal or business policies may require this service to be contained in-house and not outsourced. Determine the requirements for message hygiene.

Record this information in Table A-1 in Appendix A: “Job Aids.”

### **Availability Requirements**

Availability refers to a level of service provided by applications, services, or systems. To achieve high availability, fault-tolerance mechanisms must be implemented to minimize the impact of the failures of the components and dependencies of the service. Fault tolerance is achieved by implementing redundancy to single points of failure components. When planning for Microsoft Exchange availability, consider all components that are part of the messaging infrastructure.

Determine if high availability (automated continuous service) is required for each of the areas described in Table 1. Rate the importance of high availability for each functional area and record in the table. This information is used later to determine what the fault-tolerant requirements are in the project.

Table 1. Importance Rating for Availability Requirements

|  |  |
| --- | --- |
| **Functional area** | **High availability required** |
| Mailbox access – access to an individual mailbox |  |
| Mailbox access for external users |  |
| Internal message flow – email within the organization |  |
| External message flow – email outside of the organization |  |
| Message hygiene – antimalware and antispam |  |
| Voice mail access |  |

Note that cost and complexity increase for each section where high availability is a requirement.

## Task 2: Identify Technical Requirements

In this task, the technical requirements need to be identified to align with the business requirements identified in Task 1. Ask the business:

* Which forests and domains will leverage messaging services? In Task 1, the scope of the project for the organization was identified. Record which forests and domains will be impacted by this project. This is used in the next step to help clearly define the number of Exchange Server instances that will be implemented, as well as any additional forest/domain requirements.
* What are the physical locations and requirements? Based on the business requirements from Task 1, document all of the physical locations in the organization that are in scope, including:
* Network connectivity
* Mailbox counts
* PBX and voice IP gateway locations

### Message Hygiene Requirements

Message hygiene is an important part of an overall antispam and antimalware solution. If the organization chooses to host message hygiene services internally, products such as Microsoft Forefront® Protection 2010 for Exchange Server at <http://technet.microsoft.com/en-us/library/cc482977.aspx> are available. Additional hardware requirements for running hygiene on Exchange server roles must also be addressed. These are called out in each applicable section in the individual design steps below.

* **Mailbox servers.** Hygiene products are available that scan messages stored in databases on the Mailbox server. Consider setting up a scheduled scan to periodically check mail databases for infections.
* **Transport servers.** For Hub and Edge Transport servers, hygiene products are available that provide real-time scanning of all messages that are “in flight” throughout the organization. Consider implementing real-time scanning on all Transport servers in the organization.
* **Client Access servers.** On-demand scanning of messages is performed via Client Access servers in the organization because hygiene products access mailboxes in the same manner as messaging clients (through the Client Access server).
* **Unified Messaging (UM) servers.** UM servers do not directly interact with messages. No Exchange-specific hygiene products are recommended for UM servers.

This information will determine placement of services in later sections of this guide.

Record this information in Table A-2 in Appendix A.

## Step Summary

In this step, the business requirements, availability requirements, technical requirements, and message hygiene requirements were identified. The information gathered in this step was recorded in Tables A-1 and A-2 in Appendix A.

The information identified in this step will be used to define the instances of Exchange Server, and to design the server roles in subsequent steps.

# Step 2: Define the Instances of Exchange Server 2010

In the previous step, the scope of the messaging environment including the business requirements, availability, and technical requirements were identified. The next step is to begin the overall architectural design of the messaging system.

The tasks to be performed in this step are:

1. Determine messaging service instances.
2. Determine the Active Directory host forest, site, and domain.

## Task 1: Determine Messaging Service Instances

In this task, the number of instances of Exchange Server is determined. An *instance* of Exchange Server is defined as the number of complete Exchange organizations that will be deployed.

The minimum number of Exchange Server instances for an organization to implement is one. It is sometimes necessary to add additional Exchange Server instances. If any of the following conditions are true, additional instances may be appropriate:

* **Mergers and acquisitions.** The organization may be the product of mergers and/or acquisitions of several entities whose Active Directory forests will continue to be separate for business reasons with little or no synchronization of user account information between them. Note that this does not preclude sharing GAL data between all users in the organization.
* **Subsidiaries/business unit separation.** The organization may decide that it needs to support separate, contained messaging infrastructures for subsidiaries/business units.
* **Regulatory and/or legal requirements.** The organization may need to maintain separate, contained messaging infrastructures for regulatory and/or legal requirements. For example, the organization may have a requirement for users in one part of the organization to share their messaging system only among themselves, sending and receiving messages to all other users in the organization through other separate implementations of Exchange Server.
* **Business requirements in Step 1.** Any specific business requirements identified in Step 1 that will affect the overall design of Exchange Server instances should be considered.

**Note**   There can be only one instance of Exchange Server per Active Directory forest.

### Complex Considerations

Some organizations have multiple Active Directory forests and/or external business relationships whose business requirements call for Unified Messaging and address list consolidation. If there are multiple forests, Exchange can be deployed either as a cross-forest topology with Exchange in each forest and recipients synchronized to present a unified GAL, or as a resource forest topology with an Exchange forest and one or more user account forests (see <http://technet.microsoft.com/en-us/library/bb124734.aspx>). In these cases, Microsoft recommends contacting Microsoft Consulting Services or a certified Microsoft partner to consult on the design.

Record the number of Exchange Server instances in Table A-3 in Appendix A.

The remaining tasks and steps in the guide should be repeated for each instance of Exchange Server identified in this task.

## Task 2: Determine the Active Directory Host Forest, Site, and Domain

In this task, determine which Active Directory forest, site, and domain will be used to host this instance of Exchange Server 2010. Exchange Server has no specific requirements relative to forest, site, or domain locations.

This decision will inform subsequent decisions relative to domain controller placement.

Record this in Table A-4 in Appendix A.

## Step Summary

In this step, the number of messaging server instances was determined, based on the business requirements identified in Step 1. Also, the Active Directory domains that will host the messaging system were determined. The information gathered in this step was recorded in Tables A-3 and A-4 in Appendix A and will be used in subsequent steps to design the remaining server infrastructure.

## Additional Reading

* Planning Active Directory (Exchange Server 2010 SP1): <http://technet.microsoft.com/en-us/library/bb123715.aspx>
* FIM 2010 Technical Overview: <http://technet.microsoft.com/en-us/library/ff621362(WS.10).aspx>
* *Infrastructure Planning and Design Guide for Active Directory Domain Services:* <http://technet.microsoft.com/en-us/library/cc268216.aspx>
* Deploy Multiple Forest Topologies: <http://technet.microsoft.com/en-us/library/bb124734.aspx>

# Step 3: Design the Mailbox Server Infrastructure

In the previous step, the number of messaging server instances was determined based on the business requirements identified in Step 1. Also, the Active Directory forest, site, and domains that will host this instance of Exchange Server 2010 were determined. In this step, the Mailbox server role instances are designed. Mailbox servers store all Exchange mailboxes, which hold email messages, calendar items, contacts, and so on.

The tasks to be performed in this step are:

1. Determine role placement.
2. Plan for capacity.
3. Plan for performance.
4. Plan for fault tolerance.

## Task 1: Determine Role Placement

Based on the information gathered in Step 1 relative to mailboxes in the organization, determine where the Mailbox servers should be located, along with the corresponding Active Directory site in which the server will reside. Later sections of this guide address the question of how many servers should be placed in each location identified here. Items that affect the decision about where to locate Mailbox servers include:

* **Administrative and security requirements.** Most administrative and maintenance tasks can be performed remotely on Mailbox servers. Consider placing Mailbox servers in locations that are secure and that provide the physical infrastructure to support the services. When selecting a site, consider whether any expertise will be onsite in the event of a server failure that requires hands-on assistance.
* **Site autonomy.** Sometimes business requirements dictate that a location should have the ability to maintain full functionality in the event of a wide area network outage. A typical example would be a manufacturing facility that uses messaging systems to control the production equipment. This may lead to a requirement to have messaging servers at the manufacturing facility.
* **Fault-tolerance dependencies.** Exchange Server implements Mailbox server fault tolerance through the concept of active and passive mailbox databases. Active databases are in use, while passive copies are not in use until needed. Replicated copies of mailbox databases can be placed on multiple servers throughout the organization, allowing for the continuity of the data and service in the event of a failure affecting the active copy of the database. (Fault tolerance is covered in more detail in Task 4.) The placement of Mailbox servers, along with the impact to security and network connectivity, should be considered when deciding where to place both active and passive database copies.

Note   It is a product requirement that each Active Directory site that contains a Mailbox server must have at least one Hub Transport server and at least one Client Access server. The placement of Mailbox servers in this step directly affects the placement of Hub Transport servers in Step 5.

Determine where to place the Mailbox servers in the organization and record this information in Tables A-4 and A-5 in Appendix A.

## Task 2: Plan for Capacity

This task helps to determine *capacity planning*, which is defined as the total number of mailboxes the server can handle at one time. This is different than *performance planning*, which focuses on ensuring the Mailbox server performs at a certain level given a set number of clients. The Exchange Server product group has developed a tool that helps with the sizing of Mailbox servers and goes into much more detail than is possible in this guide. The Mailbox Server Role Requirements Calculator can be found at <http://msexchangeteam.com/archive/2009/11/09/453117.aspx>.

Exchange Server uses the concept of databases to store all messaging information. Each mailbox object is an attribute of an Active Directory security principle. Databases contain mailboxes that are associated with their Active Directory objects, which in turn store all of the individual items (email, calendars, contacts, and so on) within the messaging system. Exchange servers can host more than one database (5 in Standard Edition, 100 in Enterprise Edition). For fault tolerance, databases can be replicated to multiple servers, which are logically grouped together into Database Availability Groups (DAGs).

Capacity is based on the cumulative size of the individual messaging databases, and performance is based on the user profile/message throughput. At the end of this task the total amount of storage for the server will have been determined. This process should be repeated for each site that hosts Mailbox servers.

The total amount of storage required can be calculated in the following way:

* **Determine the individual database size.** Record how large of a database the organization is comfortable supporting. By default, there is a limit of 1,024 gigabytes (GB) for Standard Edition, which can be modified (see Modify a Database Size Limit at <http://technet.microsoft.com/en-us/library/bb232092.aspx>), and no limit for Enterprise Edition. The product group recommends keeping databases under 2 terabytes when implementing DAG high-availability support. Items that affect database size include:
* **Database white-space requirements.** This can be estimated as the average total size of messages sent per day, times the number of users.

For example, if 100 users send and receive an average of 10 megabytes (MB) of mail per day:

White space = 100 mailboxes x 10 MB per mailbox = 1 gigabyte (GB)

* **Transport dumpster size (used for replicated mailboxes only).** The Exchange Server product group recommends that the **MaxDumpsterSizePerDatabase** be set to 1.5 times the maximum message size limit in the environment. For an article detailing the changes in the dumpster, see Understanding High Availability and Site Resilience at <http://technet.microsoft.com/en-us/library/dd638137.aspx>.
* **Mailbox recoverable items time window.** By default, soft-deleted items are stored for 14 days and calendar items are stored for 120 days, or 30 GB of space. When single item recovery is enabled (disabled by default), there is an additional 1.2 percent increase in the size of the mailbox for a 14-day deleted item retention window. For calendar version logging data (enabled by default), there is an additional 5.8 percent increase in the size of the mailbox.

The storage requirements can be calculated as:

Recoverable Items Folder Size = (daily incoming/outgoing mail x average message size x deleted item retention window) + (mailbox quota size x 0.012) + (mailbox quota size x 0.058)

* **Determine log file size requirements.** Record how much space is required for log files for each database. Items that affect log file size include:
* **Log capacity.** Log capacity is defined as the total size of log files that can be stored on the Mailbox server at any given time. To determine the amount of space required, take the total amount of log files typically generated in a day and multiply by the number of days that the log files will be kept for replay. The product group recommends adding 20 percent to this total for log growth. See Understanding Mailbox Database and Log Capacity Factors at <http://technet.microsoft.com/en-us/library/ee832796.aspx> for more information.
* **Traditional backup and restore effects.** If the organization will perform traditional backups, plan for the amount of log file storage required between backups and restores. This assumes an Exchange-aware backup solution is used that flushes the log files after a successful backup.

This is calculated as:

Space required = size of log files x number of databases x days retention

* **Differential backups.** If the backup design includes weekly full backups and daily differential backups, the amount of disk free space available for the log files needs to be larger than an entire week of log file space to allow both backup and replay during restore.
* **Mailbox resiliency backup effects.** If the organization plans to use the mailbox resiliency and single item recovery features within Exchange Server as the backup infrastructure, and thus enabling continuous replication circular logging (CRCL), the product group recommends allocating three times the required daily log generation capacity. This ensures that when replication is suspended or is not functioning under normal parameters, the databases do not dismount due to truncation failures.
* **Determine the users per database.** Take the database size from above and divide by the mailbox size limit as recorded in Step 1, Task 1. This determines the total number of users per database. This will help determine the total number of databases required on the Mailbox server.
* **Determine the databases per server.** Record how many databases per single server the organization will support. There is a limit of 5 databases per server (active and passive) for Standard Edition and 100 databases per server for Enterprise Edition. Having multiple databases with mailboxes spread across them, instead of a single database containing all mailboxes, can shorten backup and restore times, as well as other tasks that affect an entire database at one time. This will help determine the number of Mailbox servers required at the location.
* **Determine total number of databases per location.** For each location that will host servers, divide the number of users to be hosted at the location by the total number of users per database as recorded in Step 1, Task 1. Consider adding buffer space to address potential future growth of the user population at the location. Round up the quotient to the nearest whole number. This is the minimum number of databases needed for the location.
* **Determine total servers per location.** For each location that will host servers, divide the number of databases per location by the databases per server. Round up the quotient to the nearest whole number. This is the minimum number of servers needed for the location.
* **Determine available open databases for fault tolerance.** After the number of servers has been calculated, it is possible that there will be unused space on the last server. To optimize the server usage as much as possible, this space will be used later in fault-tolerance planning as potential locations where passive database copies can be stored. Document the extra space available on each server by taking the total disk size and subtracting the active database space allocated for the server.

Record the physical disk storage requirements in Table A-6 in Appendix A.

## Task 3: Plan for Performance

Once the amount of needed storage has been determined, plan for the system’s overall performance. Performance planning includes identifying what is required from the server in order to have the server perform at a certain minimum level. The following assumes that the Mailbox server role will be the only Exchange Server role on the server.

### Processor Cores

According to the Exchange Server product group, all of the other server roles in Exchange Server are sized based on the number of processor cores in the Mailbox servers; therefore, this planning process is critical because it will be used as the basis for all other calculations. The product group recommends a minimum of two sockets with two cores per socket.

Performance planning centers around the number of processor megacycles (the amount of processor usage) required for the server. The number of megacycles per processor core is equal to the processor’s speed in megahertz (for example, a 2.4 gigahertz processor core has 2,400 available megacycles). Determine the number of megacycles per user, based on the estimated user profile and message activity, and multiply by the number of users on the server at peak times. The user count should include the passive copy database users because the passive copy might become the active copy at any time. This will result in the total megacycles required for the server.

Note   Due to inconsistency in how Hyperthreading reports performance, the product group recommends that Hyperthreading be disabled by default on all production Exchange servers.

To determine the overall processor requirements, several individual calculations will be performed in an order that leads to the cumulative total of the processor requirements for the server. Calculate the requirements as:

Megacycles required per mailbox = 1 megacycle per each 50 messages sent and/or received per day

This calculation gives the number of megacycles per individual mailbox, and is based on an estimate of how many messages are sent and/or received by one mailbox on a given day.

Total megacycles required = total number of mailboxes x megacycles per mailbox

The total megacycles are calculated by taking the total number of mailboxes to be hosted on the server and multiplying by the per mailbox megacycle requirements.

Processor cores = megacycle requirement / processor core megacycle capacity

The total megacycles are divided by the individual processor core megacycle capacity to determine the number of processor cores required in the server. Processor core megacycle capacity is defined as the organization’s desired processor platform speed (for example, processors running at 3 GHz) expressed in megacycles.

Megacycles per socket = number of cores per socket x megacycles per core

The required number of megacycles per socket is calculated by taking the number of processor cores per socket and multiplying by the number of megacycles required per processor core.

Processor sockets = total number of megacycles required / megacycles per socket

Finally, the number of processor sockets required is calculated by taking the total number of megacycles required for the server and dividing by the number of megacycles per socket.

For example: Assume 4,000 mailboxes that send and receive 300 messages each per day:

Total megacycles required = 4,000 mailboxes x (300 messages / 50) = 24,000

Number of cores required = 24,000 megacycles / 3,333 megacycles/processor core = 7.2, round up to 8

Divide this figure by 2 sockets (the minimum recommended). The results are 4 cores per processor.

The resultant number should then be distributed evenly across sockets for the server.

For example:

Number of cores required = 8

Sockets available on platform choice = 2

Number of cores per socket = 8 / 2 = 4

Number of cores required = 2

For the most recent processor performance capabilities, see <http://www.spec.org>.

Record the processor requirements in Table A-7 in Appendix A.

### Memory Subsystem

Use the following formula to calculate the memory (RAM) required for the server:

Memory required = 4 GB base RAM + ROUNDUP(# databases / 10) \* 2 +   
(# messages per day / 50 \* 3 MB)

**Calculate the memory required in the server by taking the base amount of 4 GB and adding memory required for each database’s Extensible Storage Engine (ESE) overhead, and then adding the required amount of cache memory (based on the number of users and the average messaging profile).**

* The product group recommends that a minimum of 4 GB of memory be installed in a Mailbox server and then memory be added to handle the database load. For each 10 databases, an additional 2 GB of memory is required to ensure the proper operation of the ESE. This should be used to determine the minimum value of memory in the server.
* Database cache memory required is based on the user count and average number of messages sent and received each day. An additional 3 MB to 30 MB of cache memory per user mailbox will be required. For each 50 messages each user sends or receives each day, add 3 MB to the memory requirements.

For example, a server that will handle 2,000 users with an average of 150 messages per user per day will require 18 GB of cache memory:

2,000 users x 9 MB per user at 150 messages per day

**The resultant figure will be the minimum amount of physical memory required in the server to achieve required performance metrics. Round this figure up to the closest amount of memory with which the server is configurable.**

A detailed description of the **actual** calculations that need to be performed can be found in the Exchange Server library on TechNet at <http://technet.microsoft.com/en-us/library/ee832789.aspx>.

Note   Recommended maximum is defined as the upper limit of viable processor and memory configurations based on price and performance. The recommended maximum configuration is a guideline. It isn't a support criterion, and it doesn't take into account the resource requirements of third-party applications that might access or be installed on the server. The recommended maximum configuration may change over time based on price changes and technology advancements.

Determine the amount of physical **memory** required for the server and record in   
Table A-7 in Appendix A.

### Disk Subsystem

Based on the information gathered in Step 1 relative to user mailboxes, estimate the performance requirements by making the following calculations around Input/Output Operations per Second (IOPS).

IOPS per mailbox = 0.060 x (number of messages sent/received per day / 50)

The product group has found that with 75 KB message sizes (a typical size message for most users), for each 50 messages per user per day estimated, 0.060 IOPS are expected.

Calculate the required IOPS for the server based on the total number of users to be hosted on the server by multiplying the number of users by the individual IOPS requirements.

For example, assume 2,000 mailboxes that send and/or receive an average of 300 messages per day, at an average message size of 75 KB:

One mailbox IOPS = 300 messages / 50 x 0.060 = 0.36

Server IOPS = 2,000 mailboxes x .36 IOPS = 720

In this example, the disk subsystem must be capable of handling at least 720 IOPS.

In addition to this disk spindle performance, the entire end-to-end throughput of the disk subsystem (controllers, bus, and so on) needs to be evaluated to ensure that each component in the disk subsystem is capable of sustaining the calculated IOPS on the server. Work with the organization’s storage vendor to ensure that both the IOPS and overall throughput can be met and that no system bottlenecks prevent the full IOPS figures to be reached.

Record the disk performance metrics (IOPS) required for the server in Table A-7 in Appendix A.

### Network Subsystem

The network subsystem in Exchange Server is based on the underlying operating system’s networking components, which are self-tuning and require no manual intervention. There is no specific product group guidance that requires more than a single network interface card (NIC). However, having only one NIC in the server represents a single point of failure in the network subsystem. It is common to leverage teaming NICs across different network switches in fault-tolerant mode to provide redundancy for the network subsystem.

Record the network requirements in Table A-7 in Appendix A.

### Virtualization

Microsoft supports running the Mailbox server role in virtualized environments, as long as the virtual environment is approved by the Windows Server Virtualization Validation Program. For details, see <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>. Important items to note include:

* Dynamically expanding disk volumes are not supported.
* Differencing mechanisms (for example, snapshots) are not supported.
* Exchange Server supports a virtual-to-logical processor ratio of no greater than 2:1. For example, if the host server has 8 physical processor cores, a maximum of 16 virtual processors can be used.

See <http://technet.microsoft.com/en-us/library/aa996719.aspx> for specific information relative to operations in a virtual environment.

If virtualization will be used in the project, record this information in Table A-7 in Appendix A.

## Task 4: Plan for Fault Tolerance

Using the information from Step 1 relative to fault-tolerance requirements, create a mapping of databases that require fault tolerance along with the locations that will host passive database copies. Ideally, a passive copy will exist in the same physical location but different physical server as the active database as that location best services the active users. Additional copies may be placed in the same location or other locations as desired for additional resiliency or disaster recovery.

Now that the number of passive database copies per location is known, the next activity is to determine which servers within each location will host the database copies. Use the information derived in Task 2 relative to unallocated database space on each server to determine where passive databases can reside. In situations where there is not enough space on existing servers, it will be necessary to implement additional servers that have the same capacity and performance to handle the additional databases.

The following figure shows an example of a fault tolerance configuration for Mailbox servers.

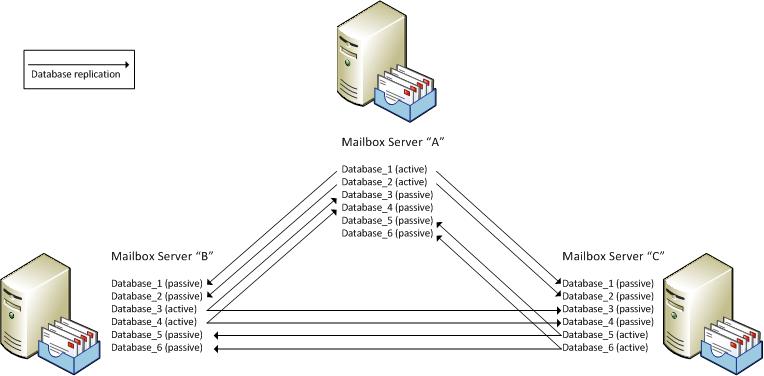


Figure 3. Example Mailbox server fault-tolerance configuration

Note   Any given server can host only one instance of a specific database.

Note   Database names must be unique across the entire organization.

Record all databases requiring fault tolerance and the locations of their passive copies in Table A-5 in Appendix A.

At this point, there should be data gathered that contains all of the Mailbox servers in the organization, what active databases each server handles, and all instances of passive databases.

## Additional Considerations

To show calculations and considerations more accurately, this guide is structured with each individual Exchange Server role separated on its own physical server platform. Included here are additional considerations the architect may need if the organization has requirements to run multiple Exchange Server roles on one server.

### Multiple Server Role Configurations

It is possible to implement multiple Exchange Server roles on a single server. In fact, all roles may exist on a single server, with the following exceptions:

* The Edge Transport server role must remain on its own separate server—it cannot be installed with any other Exchange Server role.
* Network Load Balancing (NLB) cannot be used on Mailbox servers in a database availability group because NLB cannot be combined with failover clustering.

It is important to consider the capacity and performance implications of placing multiple roles on a single server. Be sure to take the additional hardware requirements into account if the decision is made to implement multi-role servers.

## Step Summary

In this step, the Mailbox server role instances were designed. The server role placement, capacity planning, performance planning, and fault-tolerance planning were determined. The locations of all users, databases (active and passive), and servers in the organization have been recorded. The data gathered in this step was recorded in   
Tables A-4, A-5, A-6, and A-7 in Appendix A and will be used to design the infrastructure in subsequent steps.

## Additional Reading

* Exchange 2010 Mailbox Server Role Requirements Calculator: <http://msexchangeteam.com/archive/2009/11/09/453117.aspx>
* Exchange 2010 Mailbox Server Role Design Example: <http://technet.microsoft.com/en-us/library/ee832789.aspx>
* Mailbox Server Storage Design: <http://technet.microsoft.com/en-us/library/dd346703.aspx>
* Mailbox Server Processor Capacity Planning: <http://technet.microsoft.com/en-us/library/ee712771.aspx>
* Understanding Memory Configurations and Exchange Performance: <http://technet.microsoft.com/en-us/library/dd346700.aspx>
* Hardware performance metrics at the Standard Performance Evaluation Corporation: <http://www.spec.org>
* Database Copy Layout Design: <http://technet.microsoft.com/en-us/library/ff973944.aspx>
* Understanding Recoverable Items: <http://technet.microsoft.com/en-us/library/ee364755.aspx>
* Understanding Multiple Server Role Configurations in Capacity Planning: <http://technet.microsoft.com/en-us/library/dd298121.aspx>
* Windows Server Virtualization Validation Program: <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>

# Step 4: Design the Client Access Server Infrastructure

In the previous step, the design of the Mailbox server role instances for the messaging system was completed. In this step, the Client Access server roles will be designed to provide appropriate access for clients throughout the organization. The Client Access server role in Exchange Server 2010 handles all client-based communications to Mailbox servers. Protocols handled include MAPI (Microsoft Outlook®), POP3, Internet Message Access Protocol (IMAP), Microsoft Outlook Web App (OWA), and Microsoft Exchange ActiveSync®. Simple Mail Transfer Protocol (SMTP) communications are provided by Hub Transport servers, not Client Access servers.

The tasks to be performed in this step are:

1. Determine role placement.
2. Plan for capacity.
3. Plan for performance.
4. Plan for fault tolerance.

This step should be repeated for each location that contains Client Access servers.

## Task 1: Determine Role Placement

In this task, Client Access server placement will be determined. It is a product requirement that each Active Directory site that contains an Exchange Mailbox server must have at least one Client Access server. Start with one Client Access server per Active Directory site hosting Mailbox servers. Additional Client Access servers may be required, based on the remaining tasks in this step.

### Internet-Facing Client Access Server Placement

To connect to their mailboxes from outside of the organization, at least one Active Directory site must be configured to be Internet-facing for Exchange Server. If the organization will not provide client access to users outside of the firewall, this may be skipped. There is no product requirement to have more than one Internet-facing site because users will be proxied to non-Internet-facing Client Access servers through the Internet-facing site. Add at least one Client Access server for each location that will service Internet-based messaging clients.

Note   Microsoft does not support Client Access servers that are directly connected to the Internet (for example, in a perimeter network). Typically, these servers are accessed using an intermediate device (a reverse proxy) such as Microsoft Forefront Unified Access Gateway server. See <http://technet.microsoft.com/en-us/library/ee886315.aspx> for more details.

Note   All Client Access servers that will host Internet users must be in an Active Directory site that is Internet-facing.

See <http://technet.microsoft.com/en-us/library/bb310763.aspx> for more information about how messaging clients connect and communicate with Client Access servers.

Record the placement of Client Access servers in Tables A-4 and A-8 in Appendix A.

## Task 2: Plan for Capacity

This task helps to determine capacity planning, which is defined as the total number of messaging clients the server can handle at one time. Client Access servers are stateless devices that connect users to their Mailbox servers. They do not persist user data on disk; therefore, storage design is not a constraining factor. Product minimums are appropriate for disk sizing. See <http://www.microsoft.com/exchange/2010/en/us/system-requirements.aspx> for more information on Exchange Server 2010 system requirements.

## Task 3: Plan for Performance

Performance planning includes identifying what is required from the server in order to have the server perform at a certain minimum level. The following assumes that the Client Access server role will be the only Exchange Server role on the server. The Client Access server role may also be deployed on servers with other Exchange Server roles (except for the Edge Transport server role), as long as the resource requirements for the additional loads are identified and addressed.

### Processor Cores

Calculations will be performed to determine the correct number of processor cores, processor sockets, and servers to meet performance requirements. Two calculations will be performed, the first based on the ratio of mailbox processor cores to Client Access server processor cores, the second based on the number of processor cores required based on speed. The calculation with the largest number of processor cores should be used for planning Client Access servers.

The Exchange Server product group recommends planning for 3 Client Access server processor cores for every 4 Mailbox server processor cores, with a minimum of 4 processor cores and a maximum of 12 processor cores per server.

#### Calculation 1: Calculate the Number of Cores Relative to Mailbox Server Cores

For each location where Mailbox servers exist, identify how many processor cores were determined for the Mailbox servers and calculate the number of processor cores for the Client Access servers. For every 12 cores, a new Client Access server is required.

#### Calculation 2: Calculate the Number of Cores Relative to CPU Speed Requirements

Given the number of users that will be using the Client Access server, calculate estimated requirements based on server workload per user using the following guidance from the product group.

Table 2. Client Access Server Workload Cost Comparison

| **Workload** | **CPU cost (MHz/user)** |
| --- | --- |
| Outlook(MAPI) | 0.35 |
| Outlook Anywhere | 0.80 |
| Exchange ActiveSync | 1.60 |
| Exchange Web Services (Microsoft Entourage) | 0.71 |
| Outlook Web App (OWA) | 0.86 |
| IMAP4\* | 0.86 |
| POP3\* | 0.33 |
| \*IMAP4 and POP3 protocols do not support sending new mail, so the observed costs do not reflect any sent messages within the user profile. | |

Calculate the computing requirements as:

Total Client Access server computing megahertz (MHz) required = total number of client connections x CPU cost per client

To determine the total CPU megahertz requirements, take the estimated number of users in the site and multiply by the figure from Table 2 representing the client connectivity method.

For example, assume the following:

* There will be 1,400 connections in the site. Remember to take into account that users may connect from more than one client simultaneously, such as a user that also connects with a mobile device via ActiveSync.
* 750 connect via Outlook (MAPI) = 0.35 x 750 = 262.5 MHz.
* 200 connect via Outlook Anywhere = 0.80 x 200 = 160 MHz.
* 50 connect via IMAP4 = 0.86 x 50 = 43 MHz.
* 400 connect via ActiveSync = 1.60 x 400 = 640 MHz.
* Total megahertz required = 262.5 + 160 + 43 + 640 = 1105.5 MHz.

Processor cores = megahertz requirement / processor core megahertz capacity

The total megahertz are divided by the individual processor core megahertz capacity to determine the number of processor cores required in the server. Processor core megahertz capacity is defined as the organization’s desired processor platform speed (for example, processors running at 3 gigahertz).

Processor cores = greater of the two calculations

The total number of processor cores required is the greater of the number needed to satisfy Mailbox server ratios or the number required to meet CPU speed requirements.

Total sockets required = processor cores / platform capabilities

The total number of sockets required is calculated by taking the total number of processor cores and dividing by the organization’s choice of processor platform capabilities (for example, dual core or quad core).

The results are the total number of processor cores and sockets needed for the site.

Number of servers required = total sockets required / sockets per server

The total number of servers required is calculated by dividing the total number of sockets required by the organization’s choice of hardware capabilities (for example, dual- or quad-processor architectures).

Record the processor requirements in Table A-9 in Appendix A.

### Memory Subsystem

The minimum amount of memory supported by the Exchange Server product group is 4 GB overall, and the recommended amount is 2 GB of memory per processor core.

Note   Recommended maximum is defined as the upper limit of viable processor and memory configurations based on price and performance. The recommended maximum configuration is a guideline. It isn't a support criterion, and it doesn't take into account the resource requirements of third-party applications that might access or be installed on the server. The recommended maximum configuration may change over time based on price changes and technology advancements.

For example, assume the Client Access server has 4 processor cores total:

4 \* 2 GB = 8 GB physical memory recommended per Client Access server

Determine the amount of physical memory that will be implemented in the server and record in Table A-9 in Appendix A.

### Disk Subsystem

Client Access servers are devices that connect users to their Mailbox servers. They do not persist user data on disk; therefore, performance of the disk subsystem is not a constraining factor.

### Network Subsystem

The following network cost values are published from the product group and should be used to help determine the network throughput requirements for the Client Access servers.

Table 3. Client Access Server Workload Cost Comparison

| **Workload** | **Network cost (KB/sec/user)** |
| --- | --- |
| Outlook (MAPI) | 0.37 |
| Outlook Anywhere | 0.44 |
| Exchange ActiveSync (delta from Outlook) | 1.04 |
| Exchange Web Services (Microsoft Entourage) | 0.54 |
| Outlook Web App | 0.88 |
| IMAP4\* | 0.14 |
| POP3\* | 0.79 |
| \*IMAP4 and POP3 protocols do not support sending new mail (this is accomplished via SMTP through Hub Transport servers), so the observed costs do not reflect any sent messages within the user profile. | |

Calculate the requirements as:

Total Client Access server network throughput required = total number of client connections x network cost per client

To determine the total network throughput, take the estimated number of users per Client Access server and multiply by the network cost.

For example, assume the following:

* There are 1,000 users in a site.
* 750 connect via Outlook (MAPI) = 0.37 \* 750 = 277.5 kilobytes per second (KBps).
* 200 connect via Outlook Anywhere = 0.44 \* 200 = 88 KBps.
* 50 connect via IMAP4 = 0.14 \* 50 = 7 KBps.
* 400 connect via ActiveSync = 1.04 \* 400 = 416 KBps.
* Total network bandwidth required = 370 + 88 + 7 + 416 = 881 KBps.

Ensure that the NIC is capable of handling the network throughput. If the requirements exceed the capability of the current NIC, either upgrade to a faster medium or add Client Access servers to distribute the client loads.

Record the network throughput requirements in Table A-9 in Appendix A.

### Calculate the Number of Client Access Servers

To determine the total number of Client Access servers required to support capacity and performance metrics for the site, add the totals calculated in each of the subtasks above.

If the total number of Client Access servers for the site is greater than one, a Client Access server array object must be created in Active Directory and a network load balancer must be implemented. If this is not done, all client traffic within the site will connect only to the first Client Access server implemented in the site. The load balancer allows Exchange to intelligently distribute the client workload across multiples servers within the same Active Directory site. The product group recommends that a Client Access server array object be created even if there is only a single Client Access server in the Active Directory site, in the event that a second or replacement Client Access Server needs to be added. If this is not done, then manual reconfiguration of client machines may be required in order for clients to recognize the new servers.

Record the number of Client Access servers and any network load balancers in   
Table A-9 in Appendix A.

### Virtualization

Microsoft supports running the Client Access server role in virtualized environments as long as the virtual environment is approved by the Windows Server Virtualization Validation Program. For details, see <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>. Important items to note include:

* Dynamically expanding disk volumes are not supported.
* Differencing mechanisms (for example, snapshots) are not supported.
* Exchange supports a virtual-to-logical processor ratio of no greater than 2:1. For example, if the host server has 8 physical processor cores, a maximum of 16 virtual processors can be used.

See <http://technet.microsoft.com/en-us/library/aa996719.aspx> for specific information relating to operations in a virtual environment.

If hosting additional roles on this server, the performance metrics must be adjusted according to the additional loads required by those roles.

If virtualization will be used, record this in Table A-9 in Appendix A.

## Task 4: Plan for Fault Tolerance

Now that the location and configuration of each Client Access server has been established, the next task is to apply the business fault tolerance requirements from Step 1, Task 1 to determine which sites will need additional servers to be deployed and the number of servers that will be required. If the organization requires fault tolerance for mailboxes, along with Mailbox servers, Client Access servers must also be designed in for fault tolerance; without this, clients will not be able to access their mailboxes if a non-fault-tolerant Client Access server becomes unavailable.

A single Client Access server deployed in an Active Directory site that contains a Mailbox server is considered a single point of failure for that site. Wherever fault tolerance is needed, implement at least two Client Access servers in an array. If more than one Client Access server has already been identified in Task 3 above, fault tolerance will already be available as long as an array has been created. However, consider the effects on performance in the site if one server is unavailable. Consider adding an additional Client Access server to ensure minimum performance levels in the event of a single server loss.

In sites that have more than one Client Access server for performance requirements, a single additional fault-tolerant server may not be sufficient in a scenario where two or more Client Access servers become unavailable at the same time. Determine the appropriate number of additional servers that meet the organization’s fault-tolerance requirements.

After the Client Access server array has been created, a load balancer must be implemented to route internal domain-joined Outlook client traffic to the Client Access server array instead of individual servers. If this is not done, those clients will connect to the first Client Access server implemented in the site. See <http://technet.microsoft.com/en-us/library/ff625247.aspx> for further details.

Client Access server arrays are Active Directory site-bound—that is, an array that consists of Client Access servers from two different Active Directory sites cannot be implemented.

Note   In a multi-role server design, the use of network load balancers may be unsupported, depending on the combination of roles being implemented. Specifically, a Mailbox server that hosts DAGs cannot be load balanced using Network Load Balancing because Network Load Balancing is incompatible with Windows failover clustering. In this case, if Client Access server fault tolerance is required, another type of load balancer should be used or the roles should be split.

For external client connection resiliency (in the event that a single external-facing Client Access server site is unavailable), implement external-facing Client Access servers in multiple sites and configure the reverse proxy server accordingly. In the event of one site being unavailable, users will still be able to reach their messaging system via the secondary site.

The following figure shows an example of a fault tolerance configuration for Client Access servers.

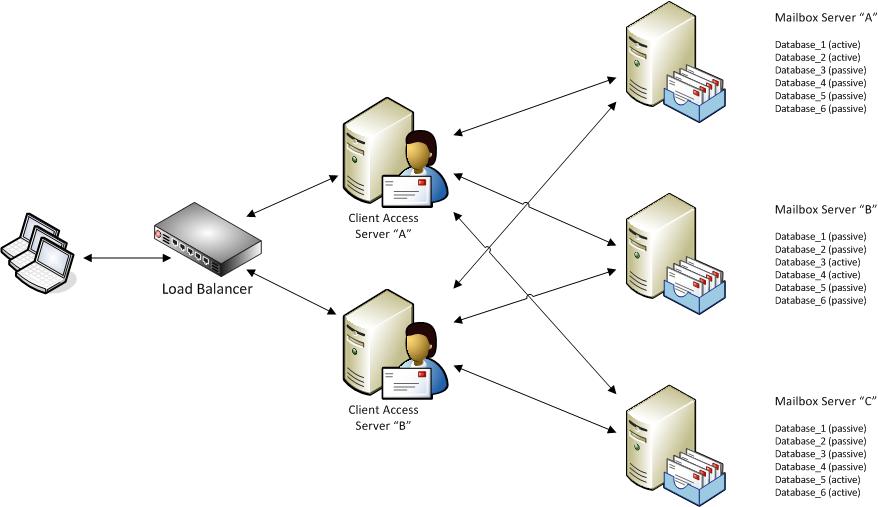


Figure 4. Example Client Access server fault-tolerance configuration

See <http://technet.microsoft.com/en-us/library/ff625247.aspx> for more details and configuration of the Client Access server role design.

For each site that will host Client Access servers, determine the number and locations of Client Access servers required to support the desired fault tolerance and record in   
Table A-8 in Appendix A.

## Step Summary

In this step, the Client Access server role instances were designed. The server role placement, capacity planning, performance planning, and fault tolerance were determined. The data gathered in this step was recorded in Tables A-4, A-8, and A-9 in Appendix A and will be used to design the infrastructure in subsequent steps.

## Additional Reading

* Understanding Load Balancing in Exchange 2010:<http://technet.microsoft.com/en-us/library/ff625247.aspx>
* Understanding Client Access:<http://technet.microsoft.com/en-us/library/bb124915.aspx>
* Understanding Processor Configurations and Exchange Performance:<http://technet.microsoft.com/en-us/library/dd346699.aspx>
* Microsoft Forefront Unified Gateway Access 2010: <http://www.microsoft.com/forefront/unified-access-gateway/en/us/default.aspx>
* How to setup an Exchange 2010 CAS Array to load balance MAPI: <http://blogs.technet.com/b/ucedsg/archive/2009/12/06/how-to-setup-an-exchange-2010-cas-array-to-load-balance-mapi.aspx>
* *Infrastructure Planning and Design Guide for Microsoft Forefront Unified Access Gateway*: <http://technet.microsoft.com/en-us/library/ee886315.aspx>
* Windows Server Virtualization Validation Program: <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>

# Step 5: Design the Hub Transport Server Infrastructure

In the previous step, the design of the Client Access server role instances for the messaging system was completed. In this step, the Hub Transport server roles will be designed. Hub Transport servers are responsible for routing all messages between all mailboxes in the organization, as well as handling SMTP messages (for example, from applications that provide email alerts).

Hub Transport servers also handle the routing of messages in and out of the organization. By default, Hub Transport servers directly connect to external servers to deliver outbound SMTP email.

Message hygiene (antivirus and antispam filtering) can be enabled on the Hub Transport servers by enabling the Edge Transport server transport agents on the Hub Transport servers. The product group recommends enabling message sanitization on Hub Transport servers to provide additional layers of message hygiene in the environment. Note that this increases the requirements of both processor and memory.

The tasks to be performed in this step are:

1. Determine role placement.
2. Plan for capacity.
3. Plan for performance.
4. Plan for fault tolerance.

This step should be repeated for each location that contains Hub Transport servers.

## Task 1: Determine Role Placement

In this task, the Hub Transport server placement is determined.

It is a product requirement that each Active Directory site that contains a Mailbox server must have at least one Hub Transport server. Start with one Hub Transport server per Active Directory site containing Mailbox servers. Additional Hub Transport servers may be required, based on the remaining tasks in this step.

Record the initial placement of Hub Transport servers in Tables A-4 and A-10 in Appendix A.

## Task 2: Plan for Capacity

In this task, the capacity requirements of the Hub Transport servers are determined.

Hub Transport servers are responsible for holding messages while in transit to Mailbox servers.

The total amount of storage required can be calculated the following way:

* **Determine message tracking space requirements.** If message tracking logs are enabled, additional capacity is required and is based on the number of messages received by the Hub Transport server. Determine the log generation rate and set a hard limit for the number of days to retain data.

For example, the product group has determined that Microsoft generates 700 MB of message tracking logs each day on each of its own Hub Transport servers, and ensures enough capacity for a week of logs, which is approximately 4.9 GB.

* **Determine protocol log space requirements.** Protocol log sizes vary depending on the activity; protocol logging is disabled by default. Microsoft generates 2.7 GB of protocol logs per day on the Hub Transport servers, and ensures that there is enough capacity for a week of logs, which is approximately 18.9 GB.
* **Determine transaction log space requirements.** Transaction logs require at least 500 MB of disk capacity for back pressure, because log creation is limited by the use of circular logging on Hub Transport servers.
* **Determine database size requirements.** The database **mail.queue** does not store items indefinitely, and the capacity reserved should be the average message size multiplied by the maximum queue to address the case when the queue is at maximum and the server is shut down.

For example, a 500,000-item queue at an average message size of 50 KB is approximately 25 GB of data in the database. Note that the amount of items will vary based on multiple factors, such as the length of time the Mailbox server is down and the amount of data transmitted over that time period.

* **Determine replicated mail database space requirements.** The transport dumpster stores in-transit messages for all replicated mailbox databases within the site until the Hub Transport server verifies that the messages have been replicated to all DAG copies. Design the Hub Transport server with enough capacity to store mail long enough for all DAGs in its site to receive replicated messages so messages can be recovered in the event of an unscheduled server outage.
* **Add buffer space.** For Hub Transport server deployments, the product group recommends adding an extra 20 percent overhead to the database size after all other factors have been considered.
* **Disk free space.** A minimum of 500 MB of free space must exist on the drive containing the message queue database, or the transport system will activate back pressure.

Storage space required = Free space requirements + message tracking requirements + protocol log requirements + transaction log requirements + database size requirements + buffer space

Total capacity required for the Hub Transport server is then calculated by adding the individual requirements above to get a sum total of disk space.

Determine the appropriate storage requirements for the Hub Transport server and record in Table A-11 in Appendix A.

## Task 3: Plan for Performance

In this task, performance planning is determined. Performance planning for Hub Transport servers involves four areas: CPU, memory, network, and disk storage performance.

The overall messaging traffic that a single Hub Transport server can handle before it becomes a potential bottleneck varies based on the organization’s particular messaging profiles and workloads. The following assumes that the Hub Transport server role will be the only Exchange Server role on the server.

### Processor Cores

Several calculations will be performed to determine the correct number of processor cores, processor sockets, and servers to meet performance requirements.

#### Calculate the Number of Cores Relative to Mailbox Server Cores

For each location where Mailbox servers exist, identify how many processor cores were determined for the Mailbox servers in Step 3, Task 3 and calculate the number of processor cores for the Hub Transport servers.

* No antivirus present: 1 Hub Transport processor core for every 7 Mailbox server processor cores.
* Antivirus present: 1 Hub Transport processor core for every 5 Mailbox server processor cores.

Total sockets required = processor cores / platform capabilities

The total number of sockets required is calculated by taking the total number of processor cores and dividing by the organization’s choice of processor platform capabilities (for example, dual-core or quad-core).

The results are the total number of processor cores and sockets needed for the site.

Number of servers required = total sockets required / sockets per server

The total number of servers required is then calculated by dividing the total number of sockets required by the organization’s choice of hardware capabilities (such as dual- or quad-processor architectures).

For example, assume the organization has a total of 16 processor cores in its mailbox servers, antivirus is implemented on the Hub Transport servers, and servers are single-socket, dual-core:

Processor cores = 16 \* (1 / 5) = 3.2, rounds up to 4 processor cores

Processor sockets = 4 processor cores / 2 cores per socket = 2 processor sockets

This calculation gives the number of processor sockets based on the number of cores distributed across the organization’s chosen platform.

Record the processor requirements in Table A-11 in Appendix A.

### Memory Subsystem

The Exchange Server product group recommends implementing a maximum of 1 GB of memory for each processor core in the server, with a 4 GB minimum. Determine the optimal memory configuration for the organization.

Note   Recommended maximum is defined as the upper limit of viable processor and memory configurations based on price and performance. The recommended maximum configuration is a guideline. It isn't a support criterion, and it doesn't take into account the resource requirements of third-party applications that might access or be installed on the server. The recommended maximum configuration may change over time based on price changes and technology advancements.

For example, assume 4 logical processor cores in the Hub Transport server:

4 \* 1 GB = 4 GB physical memory required per Hub Transport server

Determine the amount of physical memory required for the server and record in   
Table A-11 in Appendix A.

### Disk Subsystem

The Exchange Server product group has no specific guidance on performance metrics for the disk subsystem. In the absence of specific guidance, plan for complete end-to-end throughput and disk performance (IOPS) metrics based on production messaging servers in the environment that may have similar loads. This should be used as a baseline to start from, with the knowledge that the system will need to be monitored over time to ensure that it meets acceptable performance levels.

Record the disk performance metrics (IOPS) required for the server in Table A-11 in Appendix A.

### Network Subsystem

The network subsystem in Exchange Server is based on the underlying operating system’s networking components, which are self-tuning and require no manual intervention. There is no specific product group guidance that requires more than a single NIC. However, having only one NIC in the server represents a single point of failure in the network subsystem. It is common to leverage teaming NICs across different network switches in fault-tolerant mode to provide redundancy for the network subsystem.

Record the network throughput requirements in Table A-11 in Appendix A.

### Calculate the Number of Hub Transport Servers

To determine the total number of Hub Transport servers required to support capacity and performance metrics for the site, divide the number of processor sockets required by the organization’s choice of hardware platform capabilities (one socket, two sockets, and so on).

Number of Hub Transport servers = Number of processor cores / number of physical sockets / hardware platform capabilities

This calculation gives the total number of Hub Transport servers required based on distributing the number of processor cores required across available processor sockets, based on the organization’s chosen hardware platform.

For example, assume the organization has a total of 16 processor cores in its Mailbox servers, antivirus is implemented on the Hub Transport servers, and servers are single-socket, dual-core:

Processor cores = 16 \* (1 / 5) = 3.2, rounds up to 4 processor cores

Processor sockets = 4 processor cores / 2 cores per socket = 2 processor sockets

Number of servers = 2 processor sockets / 1 processor socket per server = 2 servers

Note   Hub Transport servers are fault tolerant within a site with no further configuration when more than one Hub Transport server is implemented within the site.

Record the number of Hub Transport servers in Table A-11 in Appendix A.

### Virtualization

Microsoft supports running the Hub Transport server role in virtualized environments, as long as the virtual environment is approved by the Windows Server Virtualization Validation Program. For details, see <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>. Important items to note include:

* Dynamically expanding disk volumes are not supported.
* Differencing mechanisms (for example, snapshots) are not supported.
* Exchange supports a virtual-to-logical processor ratio of no greater than 2:1. For example, if the host server has 8 physical processor cores, 16 virtual processors can be used.

See <http://technet.microsoft.com/en-us/library/aa996719.aspx> for specific information relative to operations in a virtual environment.

If hosting additional Exchange Server roles on this server, the performance metrics must be adjusted according to the additional loads required by those roles.

If virtualization will be used, record this in Table A-11 in Appendix A.

## Task 4: Plan for Fault Tolerance

The next task is to take the business fault-tolerance requirements from Step 1 and map sites where additional servers need to be deployed along with the number of servers that will be required.

A single Hub Transport server deployed in an Active Directory site is considered a single point of failure. Wherever fault tolerance is needed within a site, implement at least one additional Hub Transport server to the calculated number required above. Note that this will provide the same performance levels if one Hub Transport server becomes unavailable. If there is more than one Hub Transport server in the site, based on the amount of risk the organization is willing to take, consider adding more than one additional server to provide fault tolerance in the event that more than one server becomes unavailable at the same time.

By default, multiple Hub Transport servers are automatically load balanced within sites on the Exchange Server infrastructure. See <http://technet.microsoft.com/en-us/library/ff634392.aspx> for more information.

Note   Shadow redundancy in Exchange Server 2010 ensures that all messages arrive to their destination mailboxes properly and provides fault tolerance across the entire organization. See <http://technet.microsoft.com/en-us/library/dd351027.aspx> for more information on shadow redundancy.

The following figure shows an example of a fault tolerance configuration for Hub Transport servers.

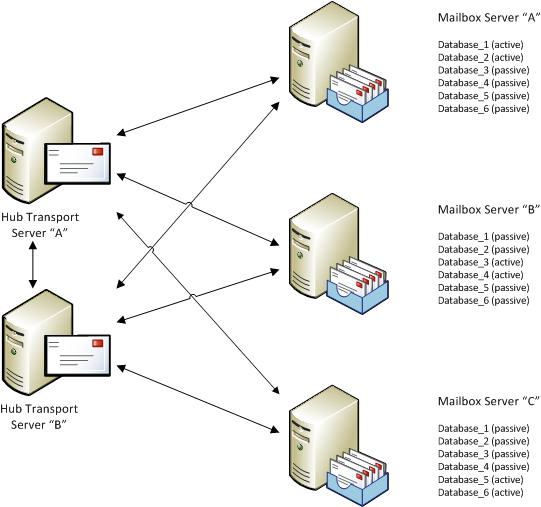


Figure 5. Example Hub Transport server fault-tolerance configuration

For each site requiring fault tolerance, record the additional server requirements in Table A-10 in Appendix A.

## Step Summary

In this step, the Hub Transport server role instances were designed. The server role placement, capacity planning, performance planning, and fault tolerance were determined. The data gathered in this step was recorded in Tables A-4, A-10 and A-11 in Appendix A and will be used to design the infrastructure in subsequent steps.

## Additional Reading

* Overview of the Hub Transport Server Role: <http://technet.microsoft.com/en-us/library/bb123494.aspx>
* Understanding Back Pressure: <http://technet.microsoft.com/en-us/library/bb201658.aspx>
* Understanding SMTP Failover and Load Balancing in Transport: <http://technet.microsoft.com/en-us/library/ff634392.aspx>
* Understanding Shadow Redundancy: <http://technet.microsoft.com/en-us/library/dd351027.aspx>
* Windows Server Virtualization Validation Program: <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>

# Step 6: Design the Edge Transport Server Infrastructure

In the previous step, the Hub Transport server roles were designed. The Edge Transport server role is optional so in this step, it will be determined if the organization requires this role. If the Edge Transport server role is not required, skip to the next step. If the Edge Transport server role is required, the remaining tasks in this step should be completed.

When deployed in an organization, the Edge Transport server handles all Internet-facing mail flow, which provides SMTP relay and smart host services for the organization. It is not required to implement an Edge Transport server (for example, if using a third-party or outsourced message relay such as Microsoft’s Forefront Online Protection for Exchange Server).

The Exchange Server product group recommends implementing the Edge Transport server role to provide a more secure external messaging experience. If the Edge Transport server role is not implemented, a solution that processes incoming messages and passes them to internal Hub Transport servers as they arrive should be implemented to provide the protection normally provided by the Edge Transport server role.

Functionally, the Edge Transport server role is very similar to the Hub Transport server role, making the individual requirements similar.

The tasks to be performed in this step are:

1. Determine role placement.
2. Plan for capacity.
3. Plan for performance.
4. Plan for fault tolerance.

## Task 1: Determine Role Placement

In this task, the Edge Transport server placement is determined.

Edge Transport server roles are deployed in the organization’s perimeter network and can be domain members or run in a workgroup. For security purposes, the product group recommends that the Edge Transport servers not be placed in the same Active Directory domain as the other Exchange Server roles.

The product group recommends that each Edge Transport server be subscribed to a single Active Directory site containing Hub Transport servers. It is possible to implement the Edge Transport server role without subscribing to Active Directory; however, key functionality (such as recipient filtering) will be lost. See <http://technet.microsoft.com/en-us/library/bb232082.aspx> for more information.

When messages arrive from outside of the organization, the Edge Transport server passes them only to Hub Transport servers in the specific Active Directory site where the Edge Transport server is subscribed.

Start with one Edge Transport server for the organization. Additional Edge Transport servers may be required, based on the remaining tasks in this step.

Consider the following when determining where to place Edge Transport servers:

* **Geographical location.** Placing Edge Transport servers in physical sites that are widely separated by geography can help control wide area network (WAN) utilization by allowing the routing of messages through servers that are closer to external users in that geography.

For example, the organization may choose to have an Edge Transport server in both the United States and Europe so that messages destined for internal users in those geographies are handled by the servers closest to them. This also has the added benefit of providing resiliency; when properly configured, if one site is down, mail can be routed automatically through the other site temporarily.

* **Network connectivity requirements.** Select locations that have acceptable external network connectivity such as bandwidth and latency.

Identify the sites and locations of Edge Transport servers and record in Tables A-4 and A-12 in Appendix A.

## Task 2: Plan for Capacity

In this task, the capacity requirements of the Edge Transport servers will be determined.

Edge Transport servers are responsible for relaying messages to internal Hub Transport servers from destinations outside of the organization.

The total amount of storage required can be calculated the following way:

* **Determine message tracking space requirements.** If message tracking logs are enabled, additional capacity is required and is based on the number of messages received by the Edge Transport server. Determine the log generation rate, and set a hard limit for the number of days to retain data.

For example, 100 MB of message tracking logs each day would require 700 MB of disk space for one week of storage.

* **Determine protocol log space requirements.** Protocol log sizes vary depending on the activity; protocol logging is disabled by default. Microsoft generates 2.7 GB of protocol logs per day on the Edge Transport servers, and ensures that there is enough capacity for a week of logs, which is approximately 18.9 GB.
* **Determine transaction log space requirements.** Transaction logs require at least 500 MB of disk capacity for back pressure, as log creation is limited by the use of circular logging on Edge Transport servers.
* **Determine database size requirements.** The database **mail.queue** does not store items indefinitely, and the capacity reserved should be the average message size multiplied by the maximum queue, to address the case when the queue is at maximum and the server is shut down.

For example, a 500,000-item queue at an average message size of 50 KB is approximately 25 GB of data in the database. Note that the amount of items will vary based on multiple factors, such as the length of time the Mailbox server is down and the amount of data transmitted over that time period.

* **Determine replicated mail database space requirements.** The transport dumpster stores in-transit messages for all replicated mailbox databases within the site until the Edge Transport server verifies that the messages have been replicated to all DAG copies. Make sure to design the Edge Transport server with enough capacity to store mail long enough for all DAGs in its site to receive replicated messages so messages can be recovered in the event of an unscheduled server outage.
* **Add buffer space.** For Edge Transport server deployments, the product group recommends adding an extra 20 percent overhead to the database size after all other factors have been considered.
* **Disk free space.** A minimum of 500 MB of free space must exist on the drive containing the message queue database, or the transport system will activate back pressure.

Storage space required = Free space requirements + message tracking requirements + protocol log requirements + transaction log requirements + database size requirements + buffer space

Total capacity required for the Edge Transport server is then calculated by adding the individual requirements above to get a sum total of disk space.

Determine the appropriate storage requirements for the Edge Transport server and record in Table A-13 in Appendix A.

## Task 3: Plan for Performance

In this task, performance planning is determined. Performance planning for Edge Transport servers involves four areas: CPU, memory, network, and disk storage performance.

The overall messaging traffic that a single Edge Transport server can handle before it becomes a potential bottleneck varies based on the organization’s particular messaging profiles and workloads. The Edge Transport server role must be the only Exchange Server role on the server because this role cannot coexist with other Exchange Server roles.

### Processor Cores

Several calculations are performed to determine the correct number of processor cores, processor sockets, and servers to meet performance requirements. Because each organization’s individual external messaging traffic is unique, estimations must be used to determine the number of processor cores.

There is no specific product group guidance on the number of cores of Edge Transport servers required; however, the product group recommends estimating the requirements by using the following metrics to set a baseline for the processor requirements on Edge Transport servers:

* Connections per second
* Messages per second
* Average message size

The number of processor cores required is based primarily on the number of connections the server will handle per second, along with the number of messages processed. This is because the server must analyze each connection and scan each accepted message. The average size of messages is also important. Monitor the processor utilization on the server to determine if the performance levels are acceptable.

See <http://technet.microsoft.com/en-us/library/dd346701.aspx> for more information on estimating these values.

Record the processor requirements in Table A-13 in Appendix A.

### Memory Subsystem

The Exchange Server product group recommends implementing a maximum of 1 GB of memory for each processor core in the server, with a 4 GB minimum:

Memory required = Number of processor cores \* 1 GB, 4 GB max

Determine the optimal memory configuration for the organization.

For example, assume 4 logical processor cores in the Edge Transport server:

4 \* 1 GB = 4 GM physical memory required per Edge Transport server

Note   Recommended maximum is defined as the upper limit of viable processor and memory configurations based on price and performance. The recommended maximum configuration is a guideline. It isn't a support criterion, and it doesn't take into account the resource requirements of third-party applications that might access or be installed on the server. The recommended maximum configuration may change over time based on price changes and technology advancements.

Determine the amount of physical memory required for the server and record in   
Table A-13 in Appendix A.

### Disk Subsystem

The Exchange Server product group has no specific guidance on performance metrics for the disk subsystem. In the absence of specific guidance, plan for complete end-to-end throughput and disk performance (IOPS) metrics based on production messaging servers in the environment that may have similar loads. This should be used as a baseline to start from, with the knowledge that the system will need to be monitored over time to ensure that it meets acceptable performance levels.

Record the disk performance metrics (IOPS) required for the server in Table A-13 in Appendix A.

### Network Subsystem

The network subsystem in Exchange Server is based on the underlying operating system’s networking components, which are self-tuning and require no manual intervention.

There is no specific product group guidance relative to network speed requirements for Edge Transport servers. As a general rule, ensure that the NIC is capable of handling the network throughput required on the server. If the requirements exceed the capability of the current NIC, either upgrade to a faster medium or add Edge Transport servers.

### Virtualization

Microsoft supports running the Edge Transport server role in virtualized environments as long as the virtual environment is approved by the Windows Server Virtualization Validation Program. For details, see <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>. Important items to note include:

* Dynamically expanding disk volumes are not supported.
* Differencing mechanisms (for example, snapshots) are not supported.
* Exchange supports a virtual-to-logical processor ratio of no greater than 2:1. For example, if the host server has 8 physical processor cores, a maximum of 16 virtual processors can be used.

See <http://technet.microsoft.com/en-us/library/aa996719.aspx> for specific information relative to operations in a virtual environment.

If virtualization will be used, record this in Table A-13 in Appendix A.

## Task 4: Plan for Fault Tolerance

In this task, the business requirements for fault tolerance identified in Step 1, Task 1 will be used to design perimeter fault tolerance. A single Edge Transport server in the organization is considered a single point of failure.

Fault tolerance for Edge Transport services requires attention to two areas:

1. **Edge Transport server failure.** If the Edge Transport server becomes unavailable, all mail flowing into the organization through that server will stop. Outbound mail will also stop, although it will be queued on the originating Hub Transport server until the Edge Transport server becomes available, or until the Hub Transport server’s queue fills.

In order to provide fault tolerance, implement additional Edge Transport servers and Round-Robin Domain Name System (RRDNS) in the perimeter network. This will allow incoming SMTP messages to be distributed across more than one Edge Transport server and provide fault tolerance if one Edge Transport server fails.

1. **Edge-connected Active Directory site becomes unavailable.** In the event that the Edge Transport server is operational but cannot communicate with the internal Hub Transport servers in its Active Directory site, inbound messages will queue on the Edge Transport server until the Hub Transport servers become available, or until back pressure levels have been reached, at which point the server will begin rejecting messages. If the Edge Transport server’s inbound queue does fill, the server will then be in the same state as number 1 above. See <http://technet.microsoft.com/en-us/library/bb201658.aspx> for more information about back pressure on Edge Transport servers.

In order to provide fault tolerance, implement an additional Edge Transport server that is subscribed to a different internal Active Directory site. Also implement RRDNS or another load-balancing mechanism in the perimeter network to provide fault tolerance on the inbound SMTP communications. This will provide an additional path for messages to move between Edge Transport and Hub Transport servers. On the Edge Transport server that is connected to the unreachable Active Directory site, messages will queue until a Hub Transport server in the affected Active Directory site becomes available or the Edge Transport server’s disk fills, whichever occurs first.

Figure 6 shows an example of a fault tolerance configuration for Edge Transport servers.

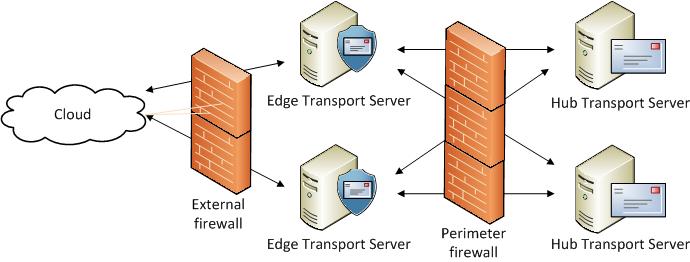


Figure 6. Example Edge Transport server fault tolerance

For each site requiring fault tolerance, record the additional server requirements in Table A-12 in Appendix A.

## Step Summary

In this step, the Edge Transport server role instances were designed. The server role placement, capacity planning, performance planning, and fault tolerance were determined. The data gathered in this step was recorded in Tables A-4, A-12 and A-13 in Appendix A.

## Additional Reading

* Overview of the Edge Transport Server Role: <http://technet.microsoft.com/en-us/library/bb124701.aspx>
* Microsoft Forefront Protection 2010 for Exchange Server: <http://technet.microsoft.com/en-us/library/cc482977.aspx>
* Forefront Online Protection for Exchange: <http://technet.microsoft.com/en-us/forefront/cc540243.aspx>
* Understanding Server Role Ratios and Exchange Performance: <http://technet.microsoft.com/en-us/library/dd346701.aspx>
* Windows Server Virtualization Validation Program: <http://windowsservercatalog.com/svvp.aspx?svvppage=svvp.htm>

# Step 7: Design the Unified Messaging Server Infrastructure

In the previous step, the Edge Transport server roles were designed. In Step 1, the requirement for the Unified Messaging (UM) server role was determined. If the organization will not be deploying unified messaging roles, skip this step. In this step, the UM server role will be designed.

The tasks to be performed in this step are:

1. Determine role placement.
2. Plan for capacity.
3. Plan for performance.
4. Plan for fault tolerance.

## Task 1: Determine Role Placement

Based on the business requirements gathered in Step 1, determine where UM servers should be located. The UM server role has specific hardware requirements relative to communication with telephony devices on the network. Specifically, the components that are required in a UM implementation include:

* **A PBX or IP PBX device.** PBX and IP PBX devices that allow multiple telephone lines to be shared by multiple users.
* **An IP gateway.** IP gateways translate circuit-based protocols from PBX devices to data or networking protocols that can communicate with messaging servers via standard Voice over IP (VoIP) protocols.

Figure 7 provides an overview of an example of a UM implementation.

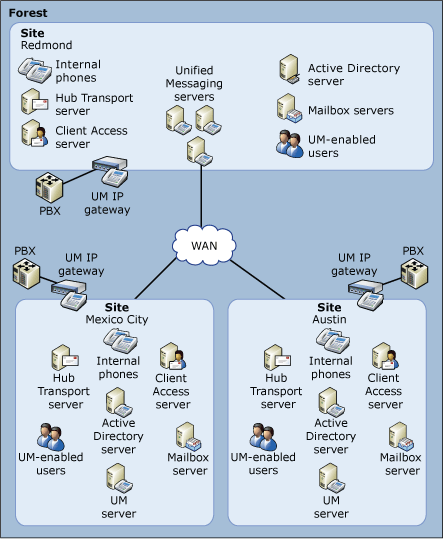


Figure 7. Overview of an example of a Unified Messaging implementation

Items that affect the decision about where to locate UM servers include:

* **Connectivity.** UM servers require reliable network connectivity between the PBX, the IP gateways, and the UM server. If there are locations that are connected to the network via links that are either too slow or have too high latency, a determination will need to be made to either eliminate the network bottleneck, or place servers locally at these locations.

The Exchange Server product group recommends that the total network round-trip time between the IP gateways and the UM servers should be less than 300ms.

* **Administrative and security requirements.** Most administrative and maintenance tasks can be performed remotely on UM servers. Consider placing UM servers in locations that are secure and that provide the physical infrastructure to support the services. When selecting a site, consider whether any expertise will be onsite in the event of a server failure that requires hands-on assistance.
* **Site autonomy.** Sometimes business requirements dictate that a location have the ability to maintain functionality in the event of a wide area network outage. A typical example would be a manufacturing facility that uses messaging systems to control the production equipment. This may lead to a requirement to have UM servers at the manufacturing facility.
* **Fault-tolerance dependencies.** Exchange Server implements UM fault tolerance by enabling call loads to be allocated across multiple servers, allowing for the continuity of the data and service in the event of a failure on the primary server. This topic is covered in more detail in later sections of this guide.

Note   There is no requirement to have UM servers in sites where Mailbox servers exist.

Note   A Hub Transport server must be available in the same Active Directory site as the UM server.

Start with one UM server and add additional servers as required to meet the above criteria. One UM server can communicate with multiple IP gateways, and multiple servers can communicate with individual IP gateways.

Determine where to place the UM servers in the organization and record this information in Tables A-4 and A-14 in Appendix A.

## Task 2: Plan for Capacity

In this task, the capacity for UM is planned. Because UM servers immediately forward all messages after they are processed to their destination mailboxes via Hub Transport servers, the disk requirements are minimal. When a UM server is unable to communicate with Hub Transport delivery servers, incoming voice messages are held on the UM server until communications with the Hub Transport servers become available. Ensure that there is enough storage to handle this scenario.

Determine the disk space requirements and record in Table A-15 in Appendix A.

## Task 3: Plan for Performance

UM servers substantially leverage the processor cores in the server, as transcoding of audio (inbound), as well as voice recognition (for voice mail previews) must be performed. Voice mail preview is particularly processor intensive; each incoming message requires approximately one minute of processor core time.

The following should be considered for performance planning:

* The product group estimates that each UM server can handle between 2,000 and 10,000 users. Plan the number of servers according to the user population.
* Capacity planning for UM servers is based on the real-time capabilities of the hardware and the number of concurrent users at a given time. UM servers automatically load-balance within their dial plan (an organization-wide Active Directory object that represents a grouping of PBXs that share common user extension numbers). Implement enough servers to handle the anticipated maximum user load expected at each PBX location.
* By default, and based on the codec selected for calls, a single UM server can accept 100 concurrent incoming calls, and a maximum of 200 concurrent voice messages. Therefore, if it’s anticipated that there will be more than this number of calls and/or voice messages, additional servers will need to be implemented.

### Processor Cores

For best performance results, the product group recommends:

* Two processor sockets.
* Sizing processor cores based on voice mail preview.
* Adding additional servers based on concurrent call volumes.

Determine the number of processor cores required and record in Table A-15 in Appendix A.

### Memory Subsystem

For best performance results, the product group recommends 2 GB of physical memory per processor core.

Note   Recommended maximum is defined as the upper limit of viable processor and memory configurations based on price and performance. The recommended maximum configuration is a guideline. It isn't a support criterion, and it doesn't take into account the resource requirements of third-party applications that might access or be installed on the server. The recommended maximum configuration may change over time based on price changes and technology advancements.

Determine the memory required and record in Table A-15 in Appendix A.

### Disk Subsystem

The product group has no specific guidance relative to the performance of the disk subsystem on UM servers. Disk access for UM messages is limited to storing and forwarding individual messages, which are not typically disk-performance sensitive. In general, the servers should be monitored for disk performance over time to ensure no disk I/O bottlenecks arise.

Determine the disk subsystem requirements for the server and record in Table A-15 in Appendix A.

### Network Subsystem

The Exchange Server network subsystem is based on the underlying operating system’s networking components, which are self-tuning and require no manual intervention.

Ensure that the NIC is capable of handling the estimated network throughput. If the requirements exceed the capability of the current NIC, either upgrade to a faster medium, or add UM servers.

The product group recommends:

* Place PBXs physically close to IP gateways.
* Place IP gateways and UM servers on the same well-connected network or within the same physical site.
* Place UM servers on the same well-connected network or within the same physical site as other computers that have Exchange Server 2010 server roles installed, including Mailbox, Hub Transport, and Client Access servers.
* End WAN connections close to where telephony equipment is located.
* In branch office scenarios or over WAN connections, use the G.723.1 codec instead of the G.711u or G.711A codec to minimize the network traffic that is passed between IP gateways and UM servers.

Determine the network requirements for the server and record in Table A-15 in Appendix A.

### Virtualization

Microsoft supports running the UM role in virtualized environments. Unified Messaging virtual machines have the following special requirements:

* Four virtual processors are required for the virtual machine. Memory should be sized using standard best practices guidance.
* Four physical processor cores are available for use at all times by each Unified Messaging role virtual machine. This requirement means that no processor oversubscription can be in use. This requirement affects the ability of the Unified Messaging role virtual machine to utilize physical processor resources.

## Task 4: Plan for Fault Tolerance

In this task, the business requirements for fault tolerance identified in Step 1, Task 1 is used to design each site requiring fault tolerance. A single UM server for the organization is considered a single point of failure.

Fault tolerance for UM servers requires attention in two areas:

1. **UM server failure.** If the UM server becomes unavailable, all voice messaging entering the organization through that server will stop until the server is brought back online.

To provide fault tolerance, implement additional UM servers. Multiple UM servers can be subscribed to the same dial plan, and will automatically fail over in the event of a single server failure. When multiple servers are added to a single dial plan, if the UM server is unavailable, the IP gateway will try to connect to the UM server again after 5 seconds. If there is no response from the UM server, the IP gateway will try to connect to the next UM server in the list in a round robin method.

1. **IP gateway failure.** If the IP gateway device fails, all voice messaging entering the organization will stop.

Consider implementing multiple IP gateways that service the same dial plan, and then subscribing the UM server to the appropriate plan.

Note   It is not possible to use network load balancers on UM servers to provide fault tolerance; only the round robin method is supported.

Determine the fault-tolerance server requirements and record in Table A-14 in Appendix A.

## Step Summary

In this step, the UM server role instances were designed. The server role placement, capacity planning, performance planning, and fault tolerance requirements were determined. The data gathered in this step was recorded in Tables A-4, A-14 and A-15 in Appendix A.

## Additional Reading

* Overview of Unified Messaging: <http://technet.microsoft.com/en-us/library/bb125141.aspx>
* Understanding Unified Messaging Server Topologies: <http://technet.microsoft.com/en-us/library/bb123925.aspx>
* Understanding Unified Messaging Call Processing: <http://technet.microsoft.com/en-us/library/bb124715.aspx>
* Understanding Unified Messaging Performance and Scalability: <http://technet.microsoft.com/en-us/library/ee332331.aspx>
* Understanding Telephony Concepts and Components: <http://technet.microsoft.com/en-us/library/bb124606.aspx>
* Understanding PBX and IP PBX Configurations: <http://technet.microsoft.com/en-us/library/bb430797.aspx>

# Step 8: Define the Active Directory Domain Services Requirements

In the previous steps, a complete implementation of Exchange Server 2010 was designed. The goal of this step is to determine any requirements for Active Directory Domain Services to support an Exchange Server 2010 deployment planned in the previous steps. Exchange Server 2010 has very specific requirements of Active Directory Domain Services that may require the organization to make adjustments to its currently deployed Active Directory Domain Services infrastructure. If these requirements are not met, Exchange Server 2010 cannot be deployed.

## Define Active Directory Domain Services Server Requirements

The location of the physical Exchange servers in the organization may drive the requirement of where to place Active Directory servers. The following server requirements must be met for Exchange Server to function properly in the environment:

* **Active Directory forest.** The Active Directory forest must be at Windows Server 2003 forest functionality mode at a minimum.
* **Active Directory schema master.** The domain controller that holds the Active Directory schema operations master role must be Windows Server 2003 with Service Pack 2 at a minimum.
* **Active Directory domain controllers.** Each domain that hosts Exchange servers must have at least one domain controller that is writable. The server must run Windows Server 2003 with Service Pack 2 at a minimum.
* **Active Directory global catalog servers.** Each site that will host Exchange servers will require at least one global catalog server; depending on the activity in the site and the fault tolerance desired, two or more global catalog servers may be required.

The Exchange Server product group recommends planning for the following:

* **32-bit platforms:** 1 global catalog server processor core for every 4 Mailbox server processor cores.
* **64-bit platforms:** 1 global catalog server processor core for every 8 Mailbox server processor cores.

Record this information in Table A-16 in Appendix A.

## Step Summary

In this step, all Active Directory dependencies were identified and recorded in Table A-16 in Appendix A.

## Additional Reading

* Planning Active Directory: <http://technet.microsoft.com/en-us/library/bb123715.aspx>
* Prepare Active Directory and Domains: <http://technet.microsoft.com/en-us/library/bb125224.aspx>
* Understanding Server Role Ratios and Exchange Performance: <http://technet.microsoft.com/en-us/library/dd346701.aspx>
* Understanding Split Permissions: <http://technet.microsoft.com/en-us/library/dd638106.aspx>

# Conclusion

This guide focused on summarizing the critical design decisions, activities, and tasks required to enable a successful design of Microsoft Exchange Server 2010. It addressed the business requirements, technical aspects, and service characteristics to complete a comprehensive review of the decision-making process. When used in conjunction with product documentation, this guide can help organizations confidently plan a Microsoft Exchange Server 2010 implementation.

# Appendix A: Job Aids

The job aids in this section can be used to record data gathered throughout the Exchange Server 2010 infrastructure planning process.

Step 1. Use Table A-1 to record the answers provided by the business to determine the capabilities and scope of the project.

Table A-1. Capabilities and Scope Requirements

| Capabilities and scope requirements | In scope (Y/N)? |
| --- | --- |
| Which of the following capabilities in Microsoft Exchange Server 2010 will be used in the environment? |  |
| Internal email services |  |
| External email services |  |
| Connectivity for external users |  |
| Public folders |  |
| Unified Communications (voice mail) services |  |
| What portion of the organization is in scope? |  |
| For the parts of the organization that are in scope, do any business isolation requirements exist? |  |
| Does the organization require an operational consolidation? |  |
| Does the organization have a requirement for a unified global address list (GAL)? |  |
| Would any business policies impact messaging usage? |  |
| Does the organization require internal message hygiene services? |  |

Step 1. Use Table A-2 to record the answers provided by the technical decision makers.

Table A-2. Technical Requirements

| Technical requirements | Results |
| --- | --- |
| Which forests and domains will leverage messaging services? |  |
| What are the physical locations and requirements? |  |
| What are the message hygiene requirements for the Mailbox, Transport, and Client Access servers? |  |

Step 2. Use Table A-3 to record the Exchange Server instances.

Table A-3. Exchange Server Instances

| Exchange Server instance #1 | Exchange Server instance #2 | Exchange Server instance #3 |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Step 2. Use Table A-4 to record the Exchange organization requirements.

Table A-4. Exchange Organization Requirements

| Object | Location |
| --- | --- |
| Which Active Directory forest will be used for Exchange Server 2010? |  |
| Which Active Directory site will be used for Exchange Server 2010? |  |
| Which Active Directory domain will be used for Exchange Server 2010? |  |
| Mailbox servers |  |
| Client Access servers |  |
| Hub Transport servers |  |
| Edge Transport servers |  |
| Unified Messaging servers |  |

Step 3. Use Table A-5 to record the Mailbox server locations, corresponding AD DS sites, active databases, fault-tolerance requirements, and passive database locations.

Table A-5. Mailbox Server Locations

| Mailbox server location | Corresponding AD DS site | Active database | Fault tolerant? (Y/N) | Passive database location |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Step 3. Use Table A-6 to record the physical disk storage requirements for the Mailbox servers.

Table A-6. Physical Disk Storage Requirements for Mailbox Servers

| Description | Requirement results |
| --- | --- |
| Database size |  |
| Log file size |  |
| Users per database |  |
| Databases per server |  |
| Total number of databases per location |  |
| Total servers per location |  |
| Number of available open databases for fault tolerance |  |

Step 3. Use Table A-7 to record the performance planning requirements for the Mailbox servers.

Table A-7. Performance Planning Requirements for Mailbox Servers

| Description | Requirement results |
| --- | --- |
| Number of processor cores |  |
| Total physical memory |  |
| Disk performance metrics (IOPS) |  |
| Network requirements |  |
| Will virtualization be used? |  |

Step 4. Use Table A-8 to record the Client Access server locations and the fault-tolerance requirements.

Table A-8. Client Access Server Locations and Fault Tolerance Requirements

| Client Access server | Fault tolerant? (Y/N) | NLB? (Y/N) |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Step 4. Use Table A-9 to record the performance planning requirements for the Client Access servers.

Table A-9. Performance Planning Requirements for Client Access Servers

| Description | Requirement results |
| --- | --- |
| Number of processor cores |  |
| Total physical memory |  |
| Network throughput requirements |  |
| Number of Client Access servers |  |
| Will virtualization be used? |  |

Step 5. Use Table A-10 to record the Hub Transport server locations and the fault-tolerance requirements.

Table A-10. Hub Transport Server Locations and Fault Tolerance Requirements

| Hub Transport server location | Fault tolerant? (Y/N) |
| --- | --- |
|  |  |
|  |  |
|  |  |

Step 5. Use Table A-11 to record the capacity and performance planning requirements for the Hub Transport servers.

Table A-11. Capacity and Performance Planning Requirements for Hub   
Transport Servers

| Description | Requirement results |
| --- | --- |
| Physical storage |  |
| Number of processor cores |  |
| Total physical memory |  |
| Disk performance metrics |  |
| Network throughput |  |
| Number of Hub Transport servers |  |
| Will virtualization be used? |  |

Step 6. Use Table A-12 to record the Edge Transport server locations and the fault-tolerance requirements.

Table A-12. Edge Transport Server Locations and Fault Tolerance Requirements

|  |  |  |
| --- | --- | --- |
| Edge Transport  server site | Active Directory  site subscription | Fault tolerant? (Y/N) |
|  |  |  |
|  |  |  |
|  |  |  |

Step 6. Use Table A-13 to record the capacity and performance planning requirements for the Edge Transport servers.

Table A-13. Capacity and Performance Planning Requirements for Edge   
Transport Servers

| Description | Requirement results |
| --- | --- |
| Physical storage |  |
| Number of processor cores |  |
| Total physical memory |  |
| Disk performance metrics |  |
| Will virtualization be used? |  |

Step 7. Use Table A-14 to record the Unified Messaging server locations and the fault-tolerance requirements.

Table A-14. Unified Messaging Server Locations and Fault Tolerance Requirements

|  |  |  |
| --- | --- | --- |
| Unified Messaging site location | Unified Messaging server location | Fault tolerant? (Y/N) |
|  |  |  |
|  |  |  |
|  |  |  |

Step 7. Use Table A-15 to record the capacity and performance planning requirements for the Unified Messaging servers.

Table A-15. Capacity and Performance Planning Requirements for Unified Messaging Servers

|  |  |
| --- | --- |
| Description | Requirement results |
| Disk requirements |  |
| Number of processor cores |  |
| Total physical memory |  |
| Disk subsystem |  |
| Network server |  |
| Will virtualization be used? | Not supported |

Step 8. Use Table A-16 to record the Active Directory sites requirements.

Table A-16. Active Directory Sites Requirements

|  |  |
| --- | --- |
| Description | Requirement results |
| Active Directory forest version |  |
| Active Directory schema master version |  |
| Number of Active Directory domain controllers |  |
| Number of Active Directory global catalog servers |  |
| 32-bit Global Catalog Server processor cores |  |
| 64-bit Global Catalog Server processor cores |  |

# Appendix B: IPD in Microsoft Operations Framework 4.0

Microsoft Operations Framework (MOF) 4.0 offers integrated best practices, principles, and activities to assist an organization in achieving reliable solutions and services. MOF provides guidance to help individuals and organizations create, operate, and support technology services, while helping to ensure the investment in technology delivers expected business value at an acceptable level of risk. MOF’s question-based guidance helps to determine what is needed for an organization now, as well as providing activities that will keep the organization running efficiently and effectively in the future.

Use MOF with IPD guides to ensure that people and process considerations are addressed when changes to an organization’s technology services are being planned.

* Use the Plan Phase to maintain focus on meeting business needs, consider business requirements and constraints, and align business strategy with the technology strategy. IPD helps to define an architecture that delivers the right solution as determined in the Plan Phase.
* Use the Deliver Phase to build solutions and deploy updated technology. In this phase, IPD helps IT pros design their technology infrastructures.
* Use the Operate Phase to plan for operations, service monitoring and control, as well as troubleshooting. The appropriate infrastructure, built with the help of IPD guides, can increase the efficiency and effectiveness of operating activities.
* Use the Manage Layer to work effectively and efficiently to make decisions that are in compliance with management objectives. The full value of sound architectural practices embodied in IPD will help deliver value to the top levels of a business.



Figure B-1. The architecture of MOF 4.0

# Appendix C: Microsoft Exchange Server 2010 in Microsoft Infrastructure Optimization

The Infrastructure Optimization (IO) Model at Microsoft groups IT processes and technologies across a continuum of organizational maturity. (For more information, see <http://www.microsoft.com/infrastructure>.) The model was developed by industry analysts, the Massachusetts Institute of Technology (MIT) Center for Information Systems Research and Microsoft's own experiences with its enterprise customers. A key goal for Microsoft in creating the Infrastructure Optimization Model was to develop a simple way to use a maturity framework that is flexible and can easily be applied as the benchmark for technical capability and business value.

IO is structured around three information technology models: Core Infrastructure Optimization, Application Platform Optimization, and Business Productivity Infrastructure Optimization. Microsoft Exchange Server 2010 gives users the power to access electronic communications securely from anywhere the end user is located from multiple devices. According to the Core Infrastructure Optimization Model, providing a properly designed messaging infrastructure will help move an organization to the Dynamic level.

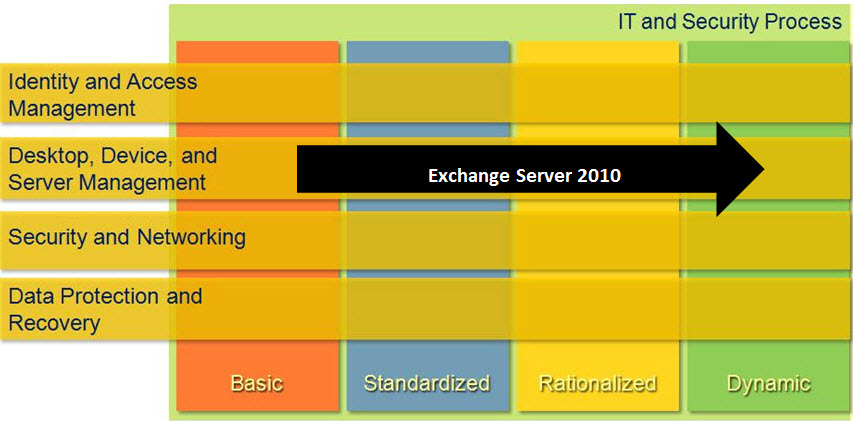


Figure C-1. Mapping of Microsoft Exchange Server 2010 technology into the Core Infrastructure Model

# Version History

|  |  |  |
| --- | --- | --- |
| **Version** | **Description** | **Date** |
| 1.1 | Minor updates in Step 2, Task 1 and Step 2, Additional Reading section.  Deleted Connectivity paragraph in Step 3, Task 1; minor updates in Step 3, Task 2.  Added a note in Memory Subsystem sections in Step 3, Task 3; Step 4, Task 3; Step 5, Task 3; Step 6, Task 3; and Step 7, Task 3.  Added verbiage in Step 7, Task 3, in the Virtualization section. | July 2011 |
| 1.0 | First release. | December 2010 |

# 

# Acknowledgments

The Solution Accelerators–Management and Infrastructure (SA-MI) team acknowledges and thanks the people who produced the *Infrastructure Planning and Design Guide for Microsoft Exchange Server 2010 with Service Pack 1*. The following people were either directly responsible for or made a substantial contribution to the writing, development, and testing of this guide.

**Contributors:**

* Jude Chosnyk – *GrandMasters*
* Karl Grunwald – *Microsoft*
* Mike Kaczmarek– *Microsoft*
* Joel Stidley – *Xtreme Consulting Group*
* Melissa Stowe – *Microsoft*
* Henry Webb – *Microsoft*

**Reviewers:**

* Rabih Andari – *Microsoft*
* Nicolas Blank – Technical Consultant
* Jon Hoerlein – *Microsoft*
* Scott Schnoll – *Microsoft*
* Tony Smith – *Microsoft*
* Serdar Soysal – *Microsoft*
* Daphne Vink – Capgemini
* Jaap Wesselius – Technical Consultant

**Editor:**

* Laurie Dunham – *Xtreme Consulting Group*

## Feedback

Please direct questions and comments about this guide to [ipdfdbk@microsoft.com](mailto:ipdfdbk@microsoft.com?subject=IPD%20-%20Microsoft%20Exchange%20Server%202010).

We value your feedback on the usefulness of this guide. Please complete the following **Solution Accelerators Satisfaction Survey**, available at: <http://go.microsoft.com/fwlink/?LinkID=132579>, and help us build better guidance and tools.