Streamlined Topologies for SharePoint Server 2013

Topology design guidance for maximizing system resources

A new approach to farm design

As an alternative to the traditional farm design, Microsoft® SharePoint® 2013 topologies can be designed to optimize system resources and to maximize performance for users.

Optimizing each tier

- *Front-end servers* — Service applications, services, and components that serve user requests directly are placed on front-end servers. These servers are optimized for fast performance.
- *Batch-processing servers* — Service applications, services, and components that process background tasks are placed on a middle-tier of servers referred to as batch processing servers. These servers are optimized to maximize system resources. These servers can tolerate greater loads because these tasks do not affect performance observed by users.
- *Database-servers* — guidance for deploying database servers remains the same.

In a small farm, server roles can be combined on one or two servers. For example, front-end services and batch-processing services can be combined on a single server or on two or more servers to achieve redundancy.

Scaling out
The front-end, batch processing, and database tiers are standardized. When another server is needed at one of these layers, an identically configured server is added.

Specialized workloads
Some service applications can cause spikes in performance, such as Excel Calculation Services or PerformancePoint. If an organization uses these service applications heavily, the recommendation is to place these on dedicated servers. If these service applications are used regularly, they can be placed on front-end servers.

Search
The search workload uses a lot of resources. When scaling beyond two batch-processing servers, place this role on dedicated servers. For more information about configuring search components, see the following model: Enterprise Search Architectures for SharePoint Server 2013.

Distributed Cache and Request Management
For small and medium-size architectures, Distributed Cache can remain on the front-end servers. Beyond 10,000 users this service is expected to work better on dedicated servers. At this scale, Request Management can be added and shared on the same servers with Distributed Cache. Request Manager is CPU intensive. Distributed Cache is memory intensive.

Server roles

Front-end servers — optimize for low latency
- Access Services
- Business Data Connectivity
- Managed Metadata
- User Profile

Distributed Cache and Request Management Servers — optimize for very high throughput

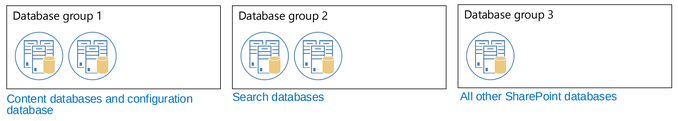
Batch-processing servers — optimize for load
- User Profile Synchronization
- Workflow
- Machine Translation
- Work Management

Specialized workloads (if needed) — optimize for medium throughput
- Search
- Excel Calculation
- PerformancePoint
- Project

Database servers — optimize for throughput

Scaling the database layer with storage groups

Storage groups
Storage groups is a concept in which similar types of databases are grouped together and scaled out independent of the rest of the databases based on need. All databases within a storage group are treated the same with backup procedures and restore protocols. The best practice is to include the configuration database with the content database group.



Example farm topologies

< 100 users

Limited deployment

One server with all roles:
- Evaluation
- Very light and simple workloads

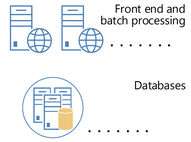
Limited deployments are typically used for product evaluation, development and testing, or for environments that have limited numbers of users and don't require fault-tolerance.

<1,000 users

Fault tolerance for simple workloads with small volumes of content

Two tiers:
- Combined front-end and batch processing servers
- Database servers

Scale the number of servers as needed.

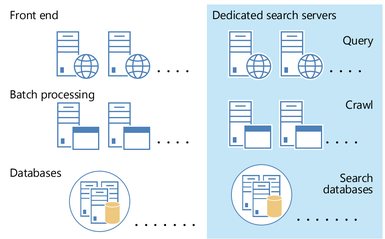


<10,000 users

Dedicated search servers for up to 10 million items.

Three tiers:
- Front-end servers
- Batch processing servers
- Database servers

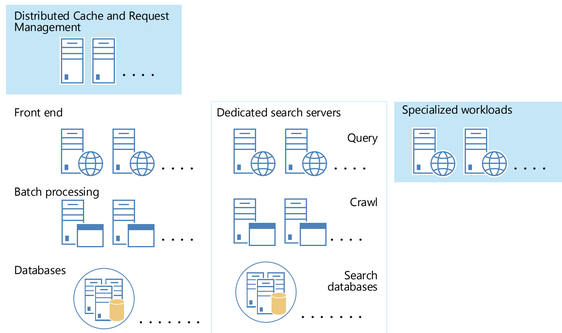
Scale the number of servers as needed.



More than 10,000 users

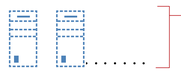
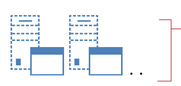
Additional server types to support large farms.

This farm represents each of the server roles that are recommended. For each server role the servers are configured identically. Scale each server role independently. Large farms benefit by adding dedicated servers for Distributed Cache and by adding Request Management.



Important: These example topologies show the progression of adding additional server roles. Servers can be virtual or physical. All numbers associated with these topologies are estimates and do not replace the need for adequate capacity planning and management.

Scale guidance for each server role

Server roles	Performance goal	Components and services		Candidates for dedicated servers
 Distributed Cache and Request Management servers	Consistent latency: • Latency — very low (<5 millisecond) • Throughput — very high • Resource utilization — medium	Distributed Cache Microsoft SharePoint Foundation Web Application Request Management		
 Front-end servers	Fast response to user requests with consistent latency: • Latency — low (<500 millisecond) • Throughput — medium • Resource utilization — low-medium	Access Services and Access Services 2010 Business Data Connectivity Central Administration Managed Metadata Microsoft SharePoint Foundation Web Application	Secure Store Service State Subscription Settings User Code User Profile Visio Graphics	Excel Calculation Performance Point Project Search Query
Batch-processing servers	Maximize resources with high throughput: • Latency — high (>1 minute) • Throughput — high • Resource utilization — high to very high	Crawl Target Machine Translation Microsoft SharePoint Foundation Web Application PowerPoint Conversion	User Profile Synchronization Word Automation Work Management Workflow timer service	Search Crawl
 Specialized workloads (if needed)	Fairly consistent latency: • Latency — low (<500 milliseconds) • Throughput — medium • Resource utilization — low-high	Excel Calculation PerformancePoint Project Search	Microsoft SharePoint Foundation Web Application	
 All databases	Fast response and consistent latency: • Latency — very low (<5 milliseconds) • Throughput — very high • Resource utilization — low-medium	For database architectures, see _____		

The Microsoft Office Division's SharePoint Server 2013 farm

A key part of the Microsoft engineering process is running a production environment using pre-release builds of SharePoint 2013. This medium-size farm supports the Microsoft Office Division.

Workload
15,000 users
2,500 unique users per hour
8,8000 active users per week
1.7 million requests per day
Collaboration, social, document management, Project
204,106 profiles
1 Web application

Dataset
1.3 Terabytes total data
1,001,141 documents
10 content databases
Largest content database—290 Gb
8,297 site collections
Largest site collection—275 Gb (tested at larger than recommended limit)

Service-level agreement (during peak hours)
The SLA is set to 99.9% availability to allow for upgrading from build to build every week during the product development cycle.

Role and hardware	Server count	Performance during peak hours: Average CPU	Memory utilization	Notes
Distributed cache and Request Management VM, 4 cores, 14 GB RAM		12%	8 GB	Two servers for availability. A load balancer is necessary to balance requests to these two servers
Front end VM, 4 cores, 14 GB RAM		45%	11 GB	Three servers allow room for spikes in performance.
Batch processing VM, 4 cores, 14 GB RAM		80%	12 GB	These servers run highly utilized to maximize the hardware. These do not process user requests.
Database 8 cores, 64 GB RAM		11%	46 GB	SQL Server is deployed to physical servers. One server is dedicated to the logging database for collecting information about the farm. Two database servers is sufficient to support the load and provide high availability.

SharePoint

Microsoft