Microsoft Skype for Business Deployment Guide

Hardware Sizing Guidelines and Capacity Planning

9/4/2025

Contents

Hardware Sizing Guidelines and Capacity Planning

This section describes the results of testing Skype for Business release 8.5.000.87, and serves as a guideline when optimizing capacity, sizing, and resource usage. Before deploying in a production environment, you must test Skype for Business in a test environment under a production load to ensure its performance meets your expectations and that it is sized properly.

Architecture Used for Testing
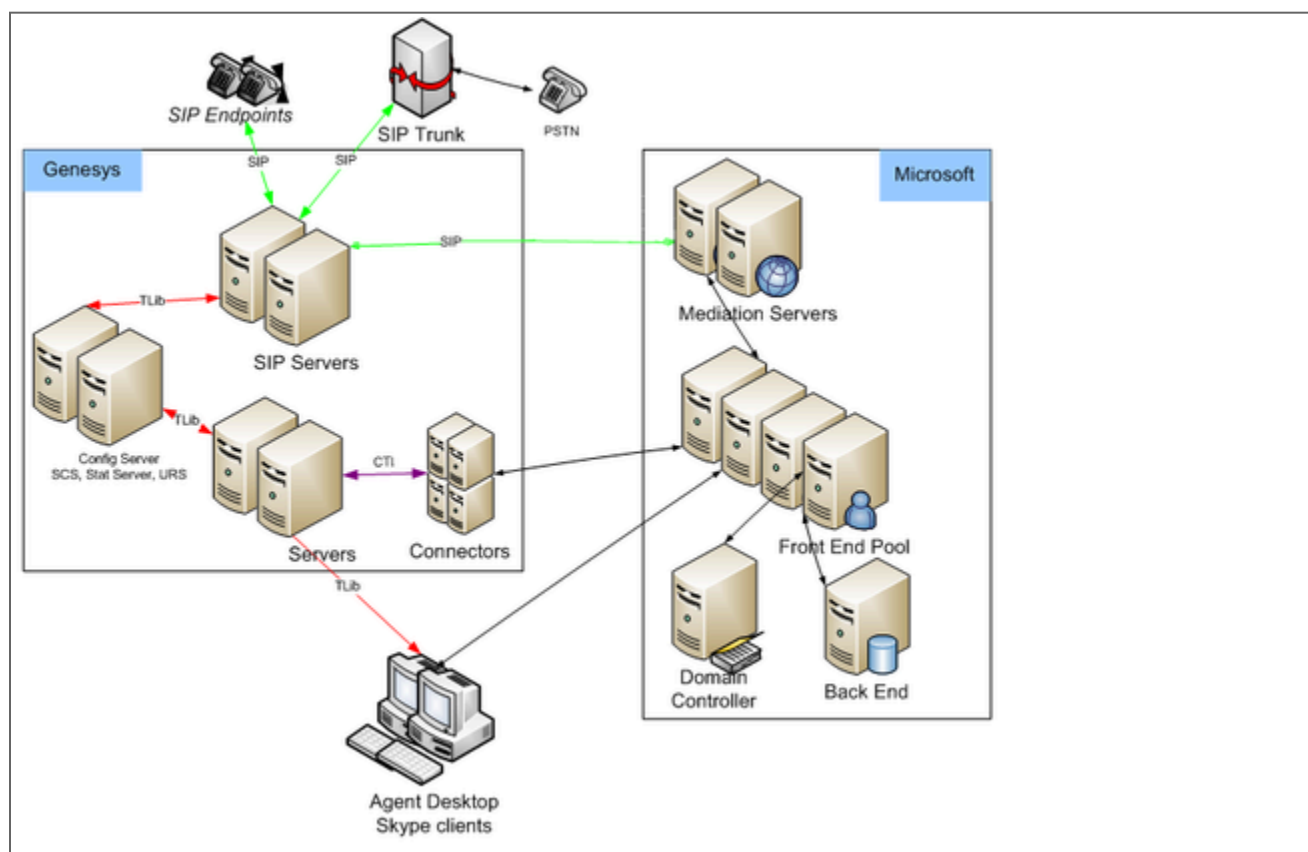
This figure depicts the architecture of the single data center Skype for Business (Microsoft) solution, and the call flow between it and the Genesys environment.

The Microsoft enterprise deployment includes:

- A pool of Mediation Servers—Translates signaling between the solution infrastructure and the PSTN gateway or SIP Trunk. The Front End Pool hosting the Genesys Contact Center applications must not be deployed using collocated Mediation Servers.

- A DNS balanced pool of Front End Servers—Shares the load. Each Front End Server maintains its own database that is synchronized with other Front End Servers.

- A Back End server and Domain Controller.

Minimum Hardware Requirements for Multimedia Connector for Skype for Business

T-Server for Skype for Business:

- 1 core CPU
- 4 GB memory

UCMA Connector:

- 6 core CPU
- 8 GB memory

See system requirements and recommendations for more information.

Test Results

Tests were performed on the following environment:

- 1 Mediation pool with 4 Mediation servers
- 1 Front End pool with 8 Front End servers
- 1 Trusted Applications pool with 8 UCMA Connectors
- 1 T-Server for Skype for Business

Genesys recommends reserving at least 2 Front End servers for HA needs.

At peak load, 1200 simultaneous calls in all scenarios were sent through the environment, and 2000 simulation agents were logged-in and engaged in the calls during the testing.

The calls-per-second rate sustained by the solution is not directly dependent on the number of deployed components or CPUs. Each new call created at the T-Server requires that you allocate resources for such activities as conference scheduling. The number of conferences created at the same time can be limited; for more information, see the Microsoft limitation. Therefore, it is not possible to achieve higher call rates, even taking into account CPU availability and load sharing among multiple components.

In general, a single instance of the T-Server with a pool of Connectors can support the following:

- Up to 1200 simultaneous calls in regular load (including at least 200 calls parked at Routing Points or processed by IVRs)

- Up to 1000 registered agents

- Up to 6 calls per second

- Up to 100 updates of attached data per call

Note that consultation calls, transfers, and conferences increase the load.