

G/On Server Sizing Reference

*Reference information for Server Sizing
and Performance Measurements*

About this document

This document gives an in-depth description of G/On performance measurement and sizing of G/On Servers.

If you do not find the information you need in this document, you may want to look in the other documents in the G/On software documentation suite:

- G/On User Guide – Getting started – Fedora
- G/On User Guide – Getting started – Windows XP
- G/On User Guide – Getting started – Windows Vista
- G/On User Guide – Getting started – Windows 7
- G/On User Guide – Getting started – Mac
- G/On User Reference
- Getting started with G/On Setup and Configuration
- Getting started with G/On Management
- Getting started with Field Deployment
- Getting started with Secure Desktop
- G/On Setup and Configuration Reference
- G/On Management Reference
- G/On Customization Reference

© Giritech A/S, 2010
Spotorno Allé 12, 2.
2630 Taastrup
Denmark
Phone +45 70.277.262

Legal Notice

Giritech reserves the right to change the information contained in this document without prior notice. Giritech® and G/On™ are trademarks and registered trademarks of Giritech A/S. Giritech A/S is a privately held company registered in Denmark. Giritech's core intellectual property currently includes the patented systems and methods known as EMCADS™. Other product names and brands used herein are the sole property of their owners. Unauthorized copying, editing, and distribution of this document is prohibited.

Contents

About this document.....	2
Contents.....	3
Introduction.....	4
G/On Versions Covered.....	4
Server OS and Database.....	4
The G/On Server Reference Hardware.....	4
Data Transmission Performance.....	5
User Load Profiles.....	5
CPU Load per User on the G/On Server Reference Hardware.....	5
CPU Load per User on Other Hardware.....	6
Max Concurrent Users	6
Weighted User Load Profiles.....	7
Authorization Performance.....	7
Scaling the Number of Cores, CPUs, or Server Machines?.....	7
Network Load.....	8
Complete Example.....	8

Introduction

This document provides information about the server hardware requirements for G/On:

- Description of the factors that affect the load on the server
- Formulae for computing the load, under given assumptions about user behavior
- Examples

Note, that G/On menu actions of type 6: Experimental RDP Connection are *not* yet covered in this document.

G/On Versions Covered

Only G/On 5.4.1 is covered by this document. Earlier versions of G/On have substantially different performance characteristics.

Server OS and Database

The information presented in this document is based on measurements with the G/On Gateway server running on Windows Server 2003. We do not expect any significant differences on Windows Server 2008.

The database was running on SQLserver on a separate machine. However, there was very little load on the database, so we do not expect any significant differences with a local database on the same machine as the Gateway Server.

The G/On Server Reference Hardware

In the document, concrete performance numbers are specified relative to the so-called *G/On Server Reference Hardware*.

The G/On Server Reference Hardware is a server with the following specs:

- Quad Core Xeon X3330 (2.66GHz, 2x3MB, 1333MHz FSB)
- 4GB (2x2GB Dual Rank DIMMs) 800MHz

Data Transmission Performance

When a user works with client-server applications, which communicate through G/On, the behavior of the user and the setup of the applications both affect the consumption of CPU on the G/On Gateway server.

OBS: G/On menu actions of type 6: Experimental RDP Connection are *not* covered in this document. The server load for this type of connections will be higher than described below, because the G/On server does a deep RDP protocol inspection for these connections.

User Load Profiles

The CPU consumption mainly depends on two characteristics of the communication between the application client and the application server:

Data Transfer Rate: the rate at which data are sent and received

Packet Frequency: The number of IP packets sent and received per time unit

These numbers can, e.g., be determined by recording the network traffic while the user is working, and then filtering it to find all the IP packets containing TCP communication between the application client and the application server.

CPU Load per User on the G/On Server Reference Hardware

Assuming a user generates data with a given Data Transfer Rate DTR and Packet Frequency PF , this user will consume the following amount of CPU resources on the Gateway server, when running on the G/On Server Reference Hardware:

$$CPU\ load\ on\ reference\ hardware = C_1 \cdot PF + C_2 \cdot DTR$$

For G/On 5.4.1, the constants have the following values:

$$C_1 = 0.01\%$$

$$C_2 = 0.0012\%$$

assuming that:

- The unit for PF is packets per second
- The unit for DTR is kilobytes per second.

Example 1

The load caused by *one* user, with a load profile of 2.5 packets/s and 34 kB/s, is computed as follows:

$$CPU\ load\ on\ reference\ hardware = 0.01\% \cdot 2.5 + 0.0012\% \cdot 34 \approx 0.066\%$$

OBS: For other server hardware and other load profiles, the CPU load will be different.

CPU Load per User on Other Hardware

The G/On Server Reference Hardware scores 58000 point in the Everest¹ Zlib performance test.

In order to estimate the CPU load on other hardware, please find out how the hardware scores in the Everest Zlib performance test, and divide this score by 58000, in order to find the relative performance index of the hardware:

$$I = \frac{\text{Everest Zlib score for chosen hardware}}{58000}$$

Now the CPU load on the chosen hardware can be computed as follows:

$$\text{CPU load} = \frac{\text{CPU load on server reference hardware}}{I}$$

Note regarding virtual machines: The procedure described here should work also for virtual server machines, provided that the Everest Zlib performance score can be measured for the virtual machine.

Example 2

Assume that we want to use an old server with an Everest Zlib performance score of only 15000. The relative performance for that server, compared to the G/On Server Reference Hardware is:

$$I = \frac{15000}{58000} \approx 0.26$$

Assume also, that we have computed the CPU load for a given user load profile on the reference hardware to be, e.g., 0.066%. The CPU load for the same user profile, on this old server will then be:

$$\text{CPU load} = \frac{0.066\%}{0.26} \approx 0.25\%$$

Max Concurrent Users

Example 3

Assume that we have computed the CPU load for a given user load profile on given server hardware to be, e.g., 0.25%. And assume that we want to keep the CPU consumption below 80%. Then the max number of concurrent users with this profile on this server hardware is:

$$\frac{80\%}{0.25\%} \approx 320.$$

¹ Everest is a commercial product from Lavalys.com. Giritech has no relationships with Lavalys. Other performance tools, which give results proportional to the Everest Zlib performance scores can be used instead of Everest Zlib.

Weighted User Load Profiles

Normally, the Data Transfer Rate and Packet Frequency will vary considerably during a realistic user work session. Variations happen, e.g, when switching between work tasks, and when there are breaks in the work, where the computer is not used.

In order to address this complexity, we recommend that one or more simple user profiles are first established, each one covering a typical work task, which is performed without interruptions.

Multiple profiles can then be combined into one, by assigning weights to the profiles, in accordance with the estimated percentage of users, who are doing each kind of task, in a given (short) period of time. The weights should be between 0 and 100% and should add up to 100%. The combined load profile can then be found by computing the weighted sum of the Data transfer Rates and the weighted sum of the Packet Frequencies.

Authorization Performance

When a user starts a G/On client and logs in, the connect and authorization process requires CPU resources on the Gateway Server, which limits how many users can log in per minute.

For G/On 5.4.1, the rule of thumb is that a server machine with a one core CPU at 2.8 GHz can handle approximately 15 log-ins per minute, provided that the CPU is not consumed by other tasks, such as data transfers from the users that have already logged in.

Server machines with two cores/CPU's at 2.8 GHz can handle max. 30 log-ins per minute.

Due to parallelization issues, only two cores/CPU's can be utilized for the connect and authorization process. So server machines with more than two cores/CPU's will only be able to handle the same number of log-ins as two core/CPU servers. Additional cores/CPU's can, however, be utilized to do data transfers for users that are already logged in.

Note: The authorization performance will be substantially improved in G/On 5.5.

Scaling the Number of Cores, CPUs, or Server Machines?

When considering *data transmission performance*, The G/On server architecture scales almost linearly, both with number of cores, number of CPUs and number of server machines.

When considering *authorization performance*, it currently does not help to add cores/CPU's beyond two per server machine, as discussed above. However, the authorization performance scales linearly in the number of server machines.

Network Load

G/On 5.4.1 adds an overhead of approximately 15% to the data transmitted. Assuming that there are N users, each behaving according to a user load profile with a given data transmission rate DTR , the load on the network can be computed as follows:

$$\text{Total Network Load} = N \cdot DTR \cdot 1.15$$

Complete Example

In this fictitious example, *Company Inc* plans to have 1200 users connecting through G/On to a Terminal Server 2003, using Remote Desktop.

Company Inc expects that the users will mainly do the following tasks through G/On:

1. Reading and responding to email.
2. Preparing power point presentations.

The G/On server hardware must be able to handle the situation where all users are working through G/On, simultaneously. *Company Inc* estimates that during the peak work periods, there will be approximately 70% users reading and responding to email, 10% preparing power point presentations, and 20% not actively using the computer.

Company Inc expects that the highest number of log-ins per minute occurs in the morning from 8:55 to 9:15, where a third of all users (i.e. 400) are expected to log in. Assuming that the log-ins are evenly distributed over this time period, *Company Inc* requires that the server machine can handle $400/20 \approx 20$ log-ins per minute.

Date Transfer load on the reference hardware

A Giritech Partner has helped *Company Inc* to make measurements of the load profiles for the two main work tasks:

Profile 1: Reading and responding to email.
PF: 2.7 packets/s, DTR: 8.9 kB/s

Profile 2: Preparing power point presentations.
PF: 2.5 packets/s, DTR: 34 kB/s

With the estimates of 70% of the users reading and responding to email while 10% prepare power point presentations, we get this weighted load profile:

Weighted Profile:
PF: $2.7 \cdot 70\% + 2.5 \cdot 10\% \approx 2.1$ packets/s
DTR: $8.9 \cdot 70\% + 34 \cdot 10\% \approx 9.6$ kB/s

The CPU load per user on the reference hardware is then:

$$\text{CPU load on reference hardware} = 0.01\% \cdot 2.1 + 0.0012\% \cdot 9.6 \approx 0.033\%$$

So for 1200 simultaneous users, the data transfer CPU load on the reference hardware will be: $1200 \cdot 0.033\% \approx 40\%$

Authorization load on the reference hardware

The reference hardware has 4 cores, so at 40% CPU load from data transfers, two of the cores will be free for simultaneously handling new log-ins at 30 users/minute. This leaves a margin, compared to the required 20 users/minute.

Network Load

With 1200 users and a data transmission rate of 9.6 kB/s:

$$\text{Total Network Load} = N \cdot \text{DTR} \cdot 1.15 = 1200 \cdot 9.6 \cdot 1.15 \simeq 13 \text{ MB/s}$$

In other words, approximately 130 Mbit/s.

Other Considerations: Fault Tolerance, Down-time, Load Fluctuations

The analysis above shows that *Company Inc* can get the desired performance by running one G/On Gateway server on a machine with specs like the G/On Server Reference Hardware.

However, it must be stressed that the analysis is based on several assumptions:

- 1) That the users only do tasks with either profile 1 or profile 2.
- 2) That the profiles adequately describe the load, when considering many users
- 3) That the worst case scenario is when 70% of the users do tasks with profile 1 and 10% do tasks with profile 2.
- 4) That the expected 400 log-ins are evenly distributed over the 20 minutes peak log-in period.

In reality, many factors can invalidate the assumptions, and it is therefore advisable to include some room for fluctuations, when sizing the server hardware.

It should also be considered whether down-time is acceptable in connection with system failures, planned maintenance and upgrades.

In order to reduce down-time and secure capacity for substantial load fluctuations, *Company Inc* decides to have two G/On Gateway servers, each running on a machine with specs like the G/On Server Reference Hardware. Even if one of these servers is down, the other has capacity enough to serve all the users. And when both machines are running, there is capacity to handle load fluctuations up to 100%.