

Capacity Management Reporting at Merrill Lynch

A Global Perspective Using TeamQuest Software

Alla Piltser, manager of the capacity planning and performance management group at Merrill Lynch, explains “capacity reconciliation.” Capacity reconciliation at Merrill Lynch is accomplished using simple management reports revealing the total available CPU resource along with the amount of that resource that is being used. Management at Merrill Lynch requested these reconciliation reports be organized in multiple ways, on a line of business or location basis, for example. The process is complicated by the fact that Merrill Lynch has a heterogeneous environment with multiple hardware and OS types.

This paper examines the various issues and concerns connected with reporting on the capacity of a distributed environment, and how to successfully compare apples with oranges using TeamQuest Performance Software.

About the Author

Alla Piltser is the manager of Distributed Capacity Planning and Performance management at Merrill Lynch. Piltser has worked with several financial institutions for the past 16 years as an independent consultant.

She has a computer science and mathematics degree from NYU and currently resides in New Jersey.

Managing the Distributed Environment: Issues and Concerns

It isn't easy monitoring a distributed environment. If server growth is left unchecked, that data center space fills up rapidly. Support for different platforms and operating systems turns provisioning and inventory management into long and laborious processes. Power supply limitations can prevent cooling upgrades or curtail the addition of servers. The end result is an increasing need to plan effectively for future capacity, but this can be very difficult without some way of knowing what is already there.

Capacity planners, therefore, are tasked with answering two key questions: How much are we using the existing environment? And what is our total installed capacity? Most of the time there isn't an easy response.

On the contrary, management wants the information boiled down to one number and delivered within a few hours.

The picture is muddled by the fact that the capacity planner is often being forced to compare apples to oranges. The sheer variety of platforms and configurations makes it difficult to draw comparisons across multi-tiered and multi-OS applications, multiple environments that are each unique (Production, Development, QA), and amid an ever-changing server population.

Clustering and failover also add to the confusion, as do a myriad of details such as hyper-threading being turned on or off, including servers with more or less cache and so on. Even if the capacity planner begins to grasp the types of hardware present, it is another matter entirely to come to grips with how that hardware is used, by what applications, and which business unit uses it.

Despite this spiraling confusion, management wants a simple answer. Top management doesn't want 15 spreadsheets full of complex equations and calculations that "show" the existing state of IT. On the contrary, management wants the information boiled down to one number and delivered within a few hours.

The Solution: Normalization

The solution to this challenge can be found via "normalization." Normalization of CPU resource capacity values allows the power of a large Unix server to be roughly compared to that of several smaller Linux servers, or for management reports to show the total horsepower of a group of servers comprised of a wide variety of

hardware and software systems.

Industry benchmark standards are useful tools for normalizing CPU processing power. SPECint_rate measurements, for example, are available for a wide range of systems and provide a common basis for comparison.

The effort to normalize can be materially assisted by leveraging TeamQuest Performance Software.

However, normalizing performance for an entire enterprise can still be a challenge, complicated by hundreds of different configurations and multi-OS applications.

The effort to normalize can be materially assisted by leveraging TeamQuest Performance Software. TeamQuest Model, for example, provides a wealth of inventory information such as model, # of CPUs and hyper-threading, and relative CPU processing power (which happens to be in the form of SPECint_rate2000 for x86 systems). Most of the servers can be easily located in TeamQuest Model's table. If a system you have cannot be found in the table, try to obtain a match at www.spec.org. At Merrill Lynch, in cases where we could only find the older SPECint_95 value, we estimated the newer benchmark by multiplying the older one by a conversion factor.

Capacity Reconciliation

Capacity reconciliation, then, is all about the construction of a monthly capacity profile for each server based on peak CPU demand, normalizing that CPU utilization, and creating zones of aggregation such as by OS, environment, line of business, application, and location.

Valuable profiles for the capacity manager to capture are monthly peak CPU utilization and CPU capacity usage for each measured server.

Valuable profiles for the capacity manager to capture are monthly peak CPU utilization and CPU capacity usage for each measured server. Ideally, these would be available for each 10-minute interval on a 24-hour timeline with approximately 30 days in each sample. By calculating the 95th percentile for all values, it is possible to derive peak demand figures for the month.

At Merrill Lynch, normalization is achieved by expressing CPU utilization in what we call CPU Capacity Units (CCUs). CCU is calculated as a function of utilization and the SPECint_rate_base2000 rating of the physical server.

We aggregate capacity data by determining the Total CPU Capacity Usage vs. Total CPU Installed Capacity in the desired aggregation

Be aware, though, that total utilization is calculated as the maximum value of the 24-hour profile for the aggregated platform.

levels or server groupings. This can be done in many different ways depending on the requirements: among all Windows servers, all Linux servers, or for all servers utilized by a line of business, for example.

We utilize a 24-hour timeline in 10-minute increments to sum up the following metrics in CCU units of measure:

All servers CPU Installed Capacity (A)

All servers CPU Capacity Usage (B)

For each 10-minute interval we calculate percentage of total utilization (UI) as follows:

$$UI = B/A * 100$$

Be aware, though, that total utilization is calculated as the maximum value of the 24-hour profile for the aggregated platform.

Note that variance accounting must be done to have the above exercise make sense in the real world. At Merrill Lynch, we noticed that our numbers were not consistent from month to month. Huge variances were showing up such as 2000 Linux servers one month and 1500 the next.

Huge variances were showing up such as 2000 Linux servers one month and 1500 the next.

This led to the realization that there was no stable base from which to compare numbers month to month. So my team had to pay attention to how many servers were modeled the previous month, calculate the number of dropped and added servers, and work out the utilization rates on any added servers. Such variances must be taken into account and included in any modeling to ensure accuracy.

End Result

The end result of all this is to highlight who really needs and who doesn't need more capacity. Using TeamQuest software, Merrill Lynch can now search by server, application and month to view the CPU Capacity profile. This can demonstrate conclusively, for example, that some users are asking for more servers even though under-utilized servers may be in their immediate vicinity.

Another factor well worth knowing is how many CCUs are available for production. While the organization as a whole might have 63,351

CCUs, 19 percent could be earmarked for QA and 14 percent for development.

In this scenario, capacity managers have to understand that only 67 percent of the total can be used for production purposes. TeamQuest can be fed that data so capacity plans are based upon effective capacity as opposed to total capacity.

TeamQuest can be fed that data so capacity plans are based upon effective capacity as opposed to total capacity.

But keep in mind that high-level summary reports don't tell the whole story. While total effective capacity may show plenty of room to breathe, it takes drilling down into each OS, line of business and application to determine that one or more of them is exceeding monthly peak CPU capacities at a particular time of day.

For example, my team discovered that we were exceeding capacity between 5:00 and 7:00 each morning on our Linux systems. By using TeamQuest to take a more detailed approach to capacity management, it is possible to spot such extremes and take them into account in the planning process.

Power of Information

The capacity manager has a wealth of information at his or her fingertips to understand installed and used capacity. This can be harnessed to optimize data center operations, control growth and become far more proactive in dealing with capacity and performance issues. Under-utilized servers and environments, for example, can be identified as candidates for virtualization and/or consolidation.

The capacity manager can also be a great friend to any area of the organization that requires more hardware. By substantiating requests to purchase new servers, the capacity manager makes it far more likely that deserving lines of business will prevail whereas wasteful units will be forced to become more efficient.

Realize, however, that capacity planning is always a work in progress. There is always something to improve and there are always changes being made — either in IT or to the various business units.

TeamQuest Corporation

www.teamquest.com

Americas

One TeamQuest Way
Clear Lake, Iowa 50428
USA
+1 641.357.2700
+1 800.551.8326
info@teamquest.com

Europe, Middle East and Africa

Box 1125
405 23 Gothenburg
Sweden
+46 (0)31 80 95 00
United Kingdom
+44 (0)1865 338031
Germany
+49 (0)69 6 77 33 466
emea@teamquest.com

Asia Pacific

Units 1001-4 10/F
China Merchants Bldg
152-155 Connaught Rd Central
Hong Kong, SAR
+852 3571-9950
asiapacific@teamquest.com

Copyright © 2008 TeamQuest Corporation
All Rights Reserved

TeamQuest and the TeamQuest logo are registered trademarks in the US, EU, and elsewhere. All other trademarks and service marks are the property of their respective owners. No use of a third-party mark is to be construed to mean such mark's owner endorses TeamQuest products or services.

The names, places and/or events used in this publication are purely fictitious and are not intended to correspond to any real individual, group, company or event. Any similarity or likeness to any real individual, company or event is purely coincidental and unintentional. NO WARRANTIES OF ANY NATURE ARE EXTENDED BY THE DOCUMENT. Any product and related material disclosed herein are only furnished pursuant and subject to the terms and conditions of a license agreement. The only warranties made, remedies given, and liability accepted by TeamQuest, if any, with respect to the products described in this document are set forth in such license agreement. TeamQuest cannot accept any financial or other responsibility that may be the result of your use of the information in this document or software material, including direct, indirect, special, or consequential damages.

You should be very careful to ensure that the use of this information and/or software material complies with the laws, rules, and regulations of the jurisdictions with respect to which it is used.

The information contained herein is subject to change without notice. Revisions may be issued to advise of such changes and/or additions. U.S. Government Rights. All documents, product and related material provided to the U.S. Government are provided and delivered subject to the commercial license rights and restrictions described in the governing license agreement. All rights not expressly granted therein are reserved.