



**The Association of System
Performance Professionals**

The **Computer Measurement Group**, commonly called **CMG**, is a not for profit, worldwide organization of data processing professionals committed to the measurement and management of computer systems. CMG members are primarily concerned with performance evaluation of existing systems to maximize performance (eg. response time, throughput, etc.) and with capacity management where planned enhancements to existing systems or the design of new systems are evaluated to find the necessary resources required to provide adequate performance at a reasonable cost.

This paper was originally published in the Proceedings of the Computer Measurement Group's 2008 International Conference.

For more information on CMG please visit <http://www.cmq.org>

Copyright 2008 by The Computer Measurement Group, Inc. All Rights Reserved

Published by The Computer Measurement Group, Inc., a non-profit Illinois membership corporation. Permission to reprint in whole or in any part may be granted for educational and scientific purposes upon written application to the Editor, CMG Headquarters, 151 Fries Mill Road, Suite 104, Turnersville, NJ 08012. Permission is hereby granted to CMG members to reproduce this publication in whole or in part solely for internal distribution with the member's organization provided the copyright notice above is set forth in full text on the title page of each item reproduced. The ideas and concepts set forth in this publication are solely those of the respective authors, and not of CMG, and CMG does not endorse, guarantee or otherwise certify any such ideas or concepts in any application or usage. Printed in the United States of America.

A KALM APPROACH TO CAPACITY PLANNING - ROAD RULES

Denise P. Kalm
CA, Inc.

Don't have time to read through manuals and books, trying to ferret out best practices for capacity planning? Learn the KPRs (Kalm Performance Rules) based on years of just doing the performance and capacity planning job, keeping my lines of business up and running well. By learning how this approach evolved, and what worked for this author, you can craft your own "best practices" and work smarter, not harder.

*"In theory, there is no difference between theory and practice; in practice, there is."
Chuck Reid*

Introduction

Let's take a journey, you and I; a journey into the future. Anyone who travels knows the value of careful trip planning, so we'll begin there. Our travels will take us from the present day – our current environment and business demands – to dates in the future to be determined by you. And we won't forget the past – the past informs the present and the future. Theoretically, you can travel as far into the future as you like (or as your data will allow), but remember,

*"Here there be tygers."
Ray Bradbury*

Pre-Trip Planning

Before starting any trip, it is helpful to know the rules of the road (axioms) that govern the journey. You would want to know whether chains are required or whether you need four-wheel drive for a road trip. For a capacity planning/performance trip, here are some of the major ones, the Seven KPRs ([Kalm Performance Rules or Kalm-isms):

1. Performance is only as good as the user thinks it is. When you assess current performance or look to the future, the data you collect on response time or throughput must be matched with data on user expectations and needs. Even if you believe the performance is good, only the user can tell you if they can get the job done with the response time offered. And since IT serves to support the business (and since we generally cannot collect real end-to-end response time numbers), user input is critical. This also means you need to have documented and clear SLAs (service level agreements.) This includes not only response time expectations, but also availability. You want them to be granular – and not based on averages. And make sure they are achievable.

And remember – response time is subjective. The number does not matter to the user. Your interest in understanding their needs and empathy for their challenges can transform an unhappy user into a satisfied one with little or no tuning performed.

2. **Empathy and respect beat out arrogance and expertise any day.** Again, this is a business-serving function. Business users may not have as much technical expertise or understanding, but they are peers. In many cases, communication skills can outweigh technical skills in dealing with the business community. Strive for a rapport with those you support; this is the unmeasured part of response time.¹ When you achieve rapport with business users, you stand to gain by really learning the business; this makes it possible to anticipate problems and do more proactive tuning. Does anyone really prefer to be paged or called at all hours?

3. **Two ears, one mouth – use in proportion.** Technical experts often find it difficult to communicate across the organization. In fairness, we really speak a different language. Anyone who travels overseas will tell you that talking more or louder does little to bridge the language barrier. Listen and ask questions – not only does this net you more information, but you look smarter and more in touch with their issues when you let the business do much of the talking. Study up on active listening – do not plan what you are going to say while they are talking. Nor should you listen with one ear, while taking in other conversations and entertaining side thoughts. Give your full attention to the conversation – there is nothing more appealing or engaging than this.

4. **Don't over-tune or over-provision.** In the 80's (and before), many shops ran a looser ship, where there was sufficient time to spend days tuning one small application and enough funding to have excess processor capacity. This is no longer true – you have to make the most of whatever resources you have on the floor. (Though one might argue that distributed servers still seem to temporarily enjoy the luxury of not running full throttle, virtualization and the tightening economy are making it essential to do a lot with less and do it faster.) Performance and capacity planning obey the 80:20 rule – you can achieve 80% of the benefits in both areas with 20% of the effort. This means you tune to achieve 80% of your desired goals and do not provision for every "worst case" imaginable, unless funding is virtually unlimited.

5. **No one ever had an "average" experience.** Averages are dangerous and rarely reflect a real person's experience with your system.² Don't rely too heavily on them. An average or "mean" is simply a mathematical construct. Look at the 90th or 95th percentile to really understand what is going on.

"Then there is the man who drowned crossing a stream with an average depth of six inches."

W.I.E. Gates

6. **Consistency is beautiful.** Remember that consistency is much more important than speed. Psychology studies have been done that prove the impact on people experiencing inconsistent response time; they are unhappy and measurably less productive. Look at the standard deviation of your response time data; if it is too large, relative to the response time number, users will not be happy.

Consider your own experience on the Net. Isn't it annoying when one page pops up quickly, but the next is very slow? Haven't you ever just gone somewhere else because the response time was erratic? Your users do too.

7. **Automate everything.** Performance and/or capacity planners are generally very senior people, having risen through the ranks of application development, systems programming or other areas. As such, your expertise and time are very valuable. Save your time for analysis and planning; let automation tools collect and sort data, produce reports and update your web site.

If you were planning a road trip, you might go to AAA to get a Trip Planner (or more likely now, to Google Maps). Given your plan, you might need to know some facts about the journey, such as places to stop for gas or for the night, conditions of the roads, or tolls along the way. The same is true for this quest; before you get started with your "journey," you will need to get some information. This will help you answer the 5 W's (who, where, when, what and why):

1. Where are you going?
2. When do you need to arrive?
3. What tools should you use?
4. Why are you doing this?
5. Who is on this journey with you?

¹ Kalm, DP "Perception is Reality – The Psychology of Performance Management,": CMG 2004 Proceedings

² Kalm, DP "The Minimum Daily Adult – The Right Metrics and The Wrong Metrics," CMG 2006 Proceedings

6. How should you make the journey?

Taken in turn, the “where” means understanding how far you want to look into the future. Though many businesses like to believe that a five-year plan makes sense, the sheer entropy of the global market doesn’t support the accuracy of such a plan. For many, an 18-month capacity plan, with quarterly estimates, is the longest “journey” possible. This allows you to sync up with budget planning efforts which are often 6 months prior to the start of a fiscal year. You really don’t have “maps” to take you much farther than that. The number of variables is too great to manage. In one case, the author’s company launched a new web site, when few companies allowed customers to really get in and work with their transaction data. Before we had advertised this site (but shortly after it went live), we had a few thousand customers hitting it. The “planning” didn’t include the idea that customers would find the site without someone pointing it out.

This example leads to the “when.” In this case, we needed to get there a lot faster than planned. When do you need to know your future capacity demands? The earlier you can start, the more information you can collect to inform your estimates. If you have little lead time, the number of assumptions you must make increases significantly. Good relations with the business translate to their desire and willingness to provide you with the information you need in a timely fashion; this can keep you from career-limiting problems.

Tools need to be in place before you start. You will need a CDB (capacity database or CMIS in ITIL v3 terms) containing historical systems data. Ideally, this data will include relationships – which business applications run on which systems applications and servers. You need some way to view and manipulate this data and you need a tool to help you do capacity planning. If you want to model performance, you will need either a queuing theory modeling tool or a simulation tool. Collect all the data, all the time; you never know what you will need.

It may seem obvious what the “why” is, but in most cases, it is a question that isn’t asked. Which applications are your loved ones, the ones that make or break your business? Which ones are growing? Where is development adding or modifying code, which could impact the path length of your transactions, affecting performance and resource utilization? What really matters to them? You should be able to state why you are on this journey very clearly, as “I am going to look at growth in my credit card application over the next 12 months, factoring in historical trends and allowing for growth based on one acquisition. No application changes are planned.” Scope is important. If you don’t know why or where, you won’t know when you get there. Finally, you should make the journey methodically and carefully, watching for speed bumps and potholes.

Travel is always more fun with company and so too is capacity planning. Even if you are the only technician charged with doing this, a few companions will make the journey easier. Among these are DBAs (to help you understand growth on databases), application programmers (to look into their future development plans), network administrators (to see if the bandwidth is there), management (to understand how much you might be able to spend if capacity upgrades are needed) and the business (to give you the projections you need for your forecasts and information on how much they make per transaction – useful to have in looking at budgeting). The application people can also help you translate business forecasts into metrics you can actually measure. Build a virtual team of experts. Among them should be those who have “made the trip” before; the well-traveled souls. They will understand how things work in your shop; there are often political or bureaucratic minefields that no class in this area can prepare you for.

Part of your planning is to understand your business. Yes, really. It is a good practice to understand each of the major applications you manage, such that when you read the paper, you can quickly estimate the probable impact of the news on your business. Knowing what the Fed was doing helped me pre-tune real estate loan applications for expected volume increases. Oddly, this kind of fore-knowledge can make you look like a genius and it really isn’t hard. To help you understand the application, meet some business users. They will be on your side once they know what you can do for them. If you can, watch the way the real end user works with

your system. User performance tuning can actually be more effective than anything you can do with system knobs.³

Next, work closely with developers to understand how they architected the application. Again, much of “performance” is already hard-coded into the application before you ever see it. As you go along, you may offer ideas of more efficient coding practices that speed response time and reduce resource demand. This is why it is important to get involved early; down the road, as code becomes more solidified, the opportunities are more limited.

If your application code is vendor-supplied, try to get involved at the initial assessment. Early tests may demonstrate that the proposed application does not scale or perhaps can never perform acceptably. If the application is already installed (and paid for), develop a relationship with the vendor. Application analysis tools may help to identify ways the vendor could improve the code, benefiting your company. This is true, even if you only have object-code-only access to the code. It is actually in their best interest to help you, as it should make ALL their customers happier; you just made it easier for them by doing the analysis.

Another important distinction to make is between foreground and background work. Whether the workload is a long-running background transaction or a daemon, you don’t want to include this work in your end user statistics. Background work – work that does not involve a user waiting for a response before entering more data – is not considered interactive. As such, this kind of work is typically looked at for throughput, rather than response time. Make sure you only include truly interactive work in your response time calculations.

Enjoying the Journey

It’s time to get on the road – are you excited yet? As with all journeys, all the planning in the world doesn’t protect you from unexpected changes and opportunities – expect them. Two Kalm-isms will help you here:

1. Trust no one. No matter what information you are given, it is only going to be a “best guess.” This is particularly true with business forecasts. Over time, you will learn who guesses high and who guesses low (and who just has no idea), and can factor accordingly.
2. Look always at the worst case scenario. Remember the age-old acronym SNAFU. Imagine what could go wrong and plan for it. Or if there is no reasonable plan, document the risk. Share this information with everyone who might care or who might later be concerned if problems do arise.

As you journey, side roads present themselves and interesting attractions advertise. The same is true with capacity planning – there are many things you could look at and way too much data; you need to pick the right data. This is where the planning helps. You already know the applications that are most important; run some quick reports to identify application patterns, so you can pick peak hours, peak days and possibly even longer term trends. An example of this is the strange phenomenon with credit card authorizations. Each Christmas season (or Hanukah or Kwanzaa), there is a huge peak in demand, based on shopping. The volume drops, but then climbs back up to this peak in July, only to reach a new peak the next holiday period. Though online shopping has made some day and hour changes to the pattern, this pattern is fairly reliable and highly useful when you have this workload. While finding peak days/hours, also notice transaction mixes, if possible. A typical “window” of interest is one hour, but you may find that five minutes or ½ hour works better for your application; this would be true for highly variable applications, where the transaction mix is not very consistent across an hour. You often must also look at various key hours, if the kinds of work being done at certain times of the day is significantly different. Another important side road is batch. Make sure that the tuning you intend to do to benefit the online workloads does not impact batch significantly enough to breach SLAs.

A well-maintained CDB is invaluable in this effort, as you can go back into history, and compare your data to the present and future forecasts. Make sure your CDB includes calculated points, such as the 90th or 95th percentile as well as averages. If you feel comfortable with statistics, include standard deviation and variance. It helps

³ Kalm, DP, “End User Performance Tuning,” CMG 2000 Proceedings

you understand the difference between an outlier or two and inconsistent performance for all. Minimum and maximum values also help you identify outliers. Get enough outliers and you need to take a side road to understand the cause.

Understand statistics, where it can help you. Statistical tools abound and help you assess the quality of your data. The bibliography contains two references on the subject of statistics to help you delve into this area, if it is new to you. Though you may not present this information to management, statistics help you make sense of data, evaluate its value and understand the end user experience.

Many travelers enjoy keeping a journal of their travels. If you are like many IT people, this will not be something you would consider, but there are two reasons to keep notes. First, you need to note any observations or concerns along the way. This is a complex process and no one can keep all of this information in easily retrievable short-term memory. Second, you need to start noting assumptions. With practice, you will begin to see where you are making assumptions; it is critical to be able to have these to hand. Every assumption is a potential mine-field (or a highway strewn with nails). One of these, if placed just right, could end your journey, invalidating your conclusions. Well-documented assumptions also give other stakeholders a chance to weigh in on the validity of these assumptions. Some may be baseline assumptions that are inherent in every study:

- Disaster recovery assessments include no more than one failing system, LPAR, server per incident
- No major resource changes will be made during the period, unless validated by the capacity management team
- No major systems programming changes will be made (new software, new releases, etc) during this period
- The systems are reasonably well tuned (you do proactive tuning, don't you?)
- The business forecasts are accurate
- CPU demand is linear with relationship to transaction volume (an assumption known not to be valid at very low or very high demand)

But others may be plan-specific:

- While transaction volume is growing, no changes are being made to the application path length that might affect response time or resource demand
- Work will be distributed to the servers in the pool based on round-robin distribution
- No mergers or acquisitions will be factored into this plan
- Growth of this application will be no more than 10% over the amount forecasted by the business
- The baseline data accurately reflects the normal transaction mix for this application
- Combining historical data from multiple days and hours gives a representative picture of resource use
- There will be no need to move additional work into this LPAR as a result of a failover or disaster

Now, to the real heart of the journey – the arcane and illuminating world of forecasting, planning and modeling. It is like exploring the Louvre; it would be impossible to really see all the art, see every room, take in everything it has to offer. So too with planning. You have the tools you have, whether you are using a spreadsheet (relatively low accuracy), linear regression (good for linearly distributed statistics), analytic queuing models or simulation models. Benchmarking is the best method, but few can afford the resources involved. None is perfect and with the complexity introduced by zOS v1r10 and the various forms of virtualization in the distributed world, modeling has become more difficult. This is where the next KPR comes in – run multiple models. Even if you are just using a very simple approach, this is still important. We will use the word “modeling” to describe whatever method you use to forecast for simplicity.

Let's look at an example. Your credit card division has decided to offer points to customers for charges, which should translate to increased processing demand, as people select this card over others. The business thinks this will increase volume by 10%. Most people would just model that and call it a day. But here is the assumption – that the business is correct. No matter how good they are at their job, this is a very uncertain

world. Running multiple models means running your plan against increasing volumes beyond 10%, perhaps 5% each time, or whatever makes sense. Keep increasing the volume till something breaks. Then, you can let the business know that if the actual increase is say 19% (instead of the forecasted 10), that they will need more CPU, disk or other resource. Or that you will need time to do some tuning to accommodate this. What does this accomplish?

In many instances, you will find that while 10% runs well, 12% may be the tipping point – the point at which response time starts missing SLAs. If you have to ramp it up to a very high volume, congratulations! You have excess capacity. But in most enterprises, this is not an issue. Note that you will also want to keep an eagle eye on the other “loved one” workloads. You can’t impact them while serving the one you are focused on. Don’t forget the ROW (rest of world).

"A man's feet should be planted in his country, but his eyes should survey the world. "
George Santayana

This kind of study opens up lines of communication with the business, as you share the results with them. They will see that you understand the challenge of their work and that you are a valuable ally. You can begin to understand how they arrived at their forecasts, which may help both of you be more accurate.

If your tools are not too cumbersome, you may wish to ask some additional questions, such as modeling the peak hours of two different applications together, if those hours are relatively close. You can try moving workloads around as some are better suited to playing nicely with others and some (like an Oracle database) run best on their own. (This is equivalent to moving children around the car on a road trip; you have to try a few things to find which ones can stand to “touch” each other and which one needs to sit in the back of the SUV with the luggage.)

As you explore the data through modeling, you are actually learning more about your applications and the systems. Note these findings; these are assumptions which over time may become axioms. You will begin to see which servers and which applications can’t push a CPU to 100% without performance impact. You will understand how important it is to see the physical side of things (100 virtual disks may just be one physical one, which will limit tuning considerably).

Finally, run your results through your intuition. Do they make sense to you? Anything that just doesn’t feel right is a sign you aren’t done yet; there are unanswered questions or unidentified assumptions. When it makes sense, then you are ready to communicate your results, request funding for resources, work with the application (or application vendor) to further tune an application or other capacity management task.

Three more KPRs before we step out of the car. You will find many more as you work in this area.

- Most single workloads don’t run well at 90+% busy on any server (but mixed ones do). Even CICS gets into dispatcher thrashing as you push the limits.
- CPU/transaction increases as utilization goes up – the cost of parallel processing. This is the “high utilization effect” based on it taking more processor time to manage the queue and dispatch increasing units of work. There is also a low utilization effect, which means CPU demand at low utilization is higher than would be expected.
- Document what you do. Then share it. Otherwise, no one else will really appreciate what you have done. And you need the feedback to improve your process. Communication builds bridges for a long-term relationship, which will benefit your capacity plan, your company, and most importantly, your reputation.

Back at Home – Reviewing the Photos

Very few capacity planners ever go back and check the accuracy of their forecasts, but this means, each time, you are starting over as a new traveler, most likely to make the same mistakes and take the same wrong turns.

"Those who cannot remember the past are condemned to repeat it."

George Santayana

Just as it can be pleasant and valuable to review vacation photos and tell stories, making note of what worked well (what you should have packed and shouldn't have, what could have made it all easier), it can also be helpful in capacity planning. In most cases, what you are doing is continuous improvement, a key part of ITIL v3. You "tune" your capacity planning methodology.

If your model did not accurately forecast what happened, where was the error? Were the forecasts too ambitious or not insightful enough? Did you miss a workload that actually ended up impacting all the others? Were you missing key data in your CDB? When you analyze your results against actuals EVERY time you go through the process, deeper learning about your system, your applications and the people involved is possible. One place where results can be correct, but stakeholders are disappointed is when you failed to manage expectations. This is where your summary of assumptions, your regular communications and interaction with all users is so important. In education, the mantra is "Tell them what you are going to tell them, tell them, then tell them what you told them." The rule of three in communication helps ensure that everyone gets it. Particularly when you are running close to capacity, or there are potential serious business risks involved, everyone involved needs to understand the situation. You will develop some practices and create some lists over time that will help render your plans more accurate each time:

1. Create a list of stakeholders. Initially, consider your management team, the application involved, and the business users. As you progress, add people like Procurement, Operations, etc., as you need them.
2. Create a list of steps to take each time, so you can expedite the process.
3. Create a running list of assumptions, crossing off the ones that did not work for you
4. Develop a communication vehicle that ensures everyone on your stakeholder list is aware of your efforts and the results. (Note: a web site, while useful, is not enough. You must drive traffic to that site – you have to tell them it is there and summarize what they will find). And communication is two-way. Find out how the business wants to communicate with you, regarding forecasts or other issues.
5. Understand the difference between "perfect" and "good enough." No plan will actually play out exactly as you wish. Determine the level of error you can live with.
6. Create your own ROTs (rules of thumb). One useful thing in this area is to create sub-SLAs. These might be response time "triggers" that notify you proactively so you can tune before users are aware of the problem, using numbers lower than the SLA values. Change in CPU demand is another metric to watch. If the rate of change is increasing (or even decreasing), this may signify a capacity event worth investigating.
7. Document all instances where capacity planning might be useful, so you can develop a regular practice of doing it.
8. Automate anything you can, leaving you time for analysis. Computers cannot do the kind of analysis and reasoning a capacity planner must.

Taking the time to evaluate each forecast will help you build a more robust system for planning, as well as build your confidence in your results. Confidence communicates. The more you believe in what you are saying, the more you will become the business' trusted advisor.

In time, you may find that you will be close enough to the business to start importing dollars into your plan. More likely the subject of another paper (or many others that already exist), a report on all aspects of their business, including the cost per transaction weighed against the profit they make (supplied by them) is an invaluable way to bridge the business-IT gap.

In the end, capacity planning is no more than providing well-performing system services to support business applications at an affordable price.

Summary

Unlike your typical vacation travel, the capacity planning journey is like the journey of life; it is a never-ending trek that involves many legs.

"A journey of a thousand miles begins with a single step."

Lao-tzu

But any journey can be helped by review of travel guides, so this paper represents a "travel guide" for capacity planning. Tips, techniques and "the stuff that works" are designed to help you not get sidetracked by "the largest ball of twine" or other wrong turns. It should also help you to "fast track" your first capacity planning exercise. As you continue, consult the CMG web site for other great articles to go deeper into specific capacity planning disciplines.

But remember, as Field Marshall von Moltke noted "no battle plan ever survives contact with the enemy," or in our case, the real world. Use these as a basis for your plan, but be prepared to adjust and course-correct as needed. Rules were just made to be broken.

"Hell, there are no rules here --- we're trying to accomplish something."

Thomas A. Edison, inventor

"If you obey all the rules, you miss all the fun."

Katherine Hepburn, actress and independent spirit

"It's a good idea to obey all the rules when you're young, just so you'll have the strength to break them when you're old."

Mark Twain, author

Bibliography

[Arnold] D, "Capacity Planning – The Practical and Political Side," CMG 2001

[Bereznay] F. Bereznay, "Did Something Change? Using Statistical Techniques to Interpret Service and Resource Metrics," CMG 2006

[Kalm] D. Kalm, "The Minimum Daily Adult – the Right Metrics and the Wrong Metrics," CMG 2006

[Kalm] D. Kalm "Perception is Reality – The Psychology of Performance Management," CMG 2004

[Kalm] D. Kalm, "The Stork Correlation – Use & Abuse of Statistics in Performance and Capacity Planning," CMG 2002

[King] G. King, "Running IBM System z at High CPU Utilization," ibm.com/supoort/techdocs, November 2, 2007