# z/VM Capacity Planning Overview

Bill Bitner
z/VM Development Lab Customer Focus and Care
bitnerb@us.ibm.com

# Trademarks

**The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.**

| | | | | | |
|---|---|---|---|---|---|
| BladeCenter* | FICON* | Performance Toolkit for VM | Storwize* | System z10* | zSecure |
| DB2* | GDPS* | Power* | System Storage* | Tivoli* | z/VM* |
| DS6000* | HiperSockets | PowerVM | System x* | zEnterprise* | |
| DS8000* | HyperSwap | PR/SM | System z* | z/OS* | |
| ECKD | OMEGAMON* | RACF* | System z9* | | |

* Registered trademarks of IBM Corporation

## The following are trademarks or registered trademarks of other companies.

Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
IT Infrastructure Library is a registered trademark of the Central Computer and Telecommunications Agency which is now part of the Office of Government Commerce.
ITIL is a registered trademark, and a registered community trademark of the Office of Government Commerce, and is registered in the U.S. Patent and Trademark Office.
Java and all Java based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
OpenStack is a trademark of OpenStack LLC. The OpenStack trademark policy is available on the OpenStack website.
TEALEAF is a registered trademark of Tealeaf, an IBM Company.
Windows Server and the Windows logo are trademarks of the Microsoft group of countries.
Worklight is a trademark or registered trademark of Worklight, an IBM Company.
UNIX is a registered trademark of The Open Group in the United States and other countries.

* Other product and service names might be trademarks of IBM or other companies.

**Notes**:

Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon considerations such as the amount of multiprogramming in the user's job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve throughput improvements equivalent to the performance ratios stated here.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.

This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This information provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs) ("SEs"). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”). No other workload processing is authorized for execution on an SE. IBM offers SE at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

# Notice Regarding Specialty Engines (e.g., zIIPs, zAAPs and IFLs):

Any information contained in this document regarding Specialty Engines ("SEs") and SE eligible workloads provides only general descriptions of the types and portions of workloads that are eligible for execution on Specialty Engines (e.g., zIIPs, zAAPs, and IFLs). IBM authorizes customers to use IBM SE only to execute the processing of Eligible Workloads of specific Programs expressly authorized by IBM as specified in the “Authorized Use Table for IBM Machines” provided at www.ibm.com/systems/support/machine_warranties/machine_code/aut.html (“AUT”).

No other workload processing is authorized for execution on an SE.

IBM offers SEs at a lower price than General Processors/Central Processors because customers are authorized to use SEs only to process certain types and/or amounts of workloads as specified by IBM in the AUT.

# Introduction

- Objectives:
  - Provoke thought and ultimately action about Capacity Planning
  - Concepts and approaches will be covered, not so much the mechanics
- Companion Piece:
  - z/VM Performance Metrics – also available from the author
- Time permitting – dialogue on what can IBM do to help in this space?



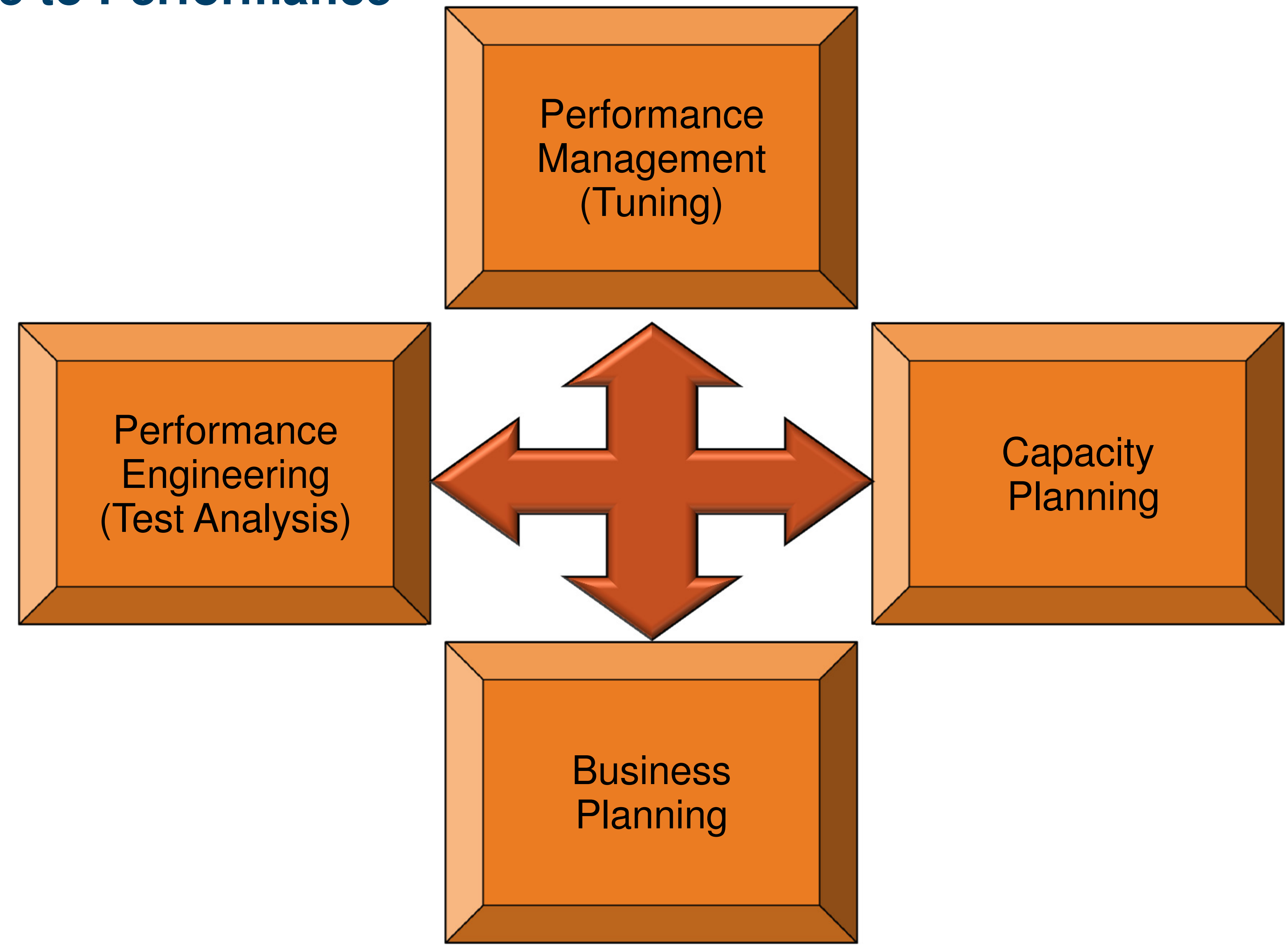
© 2013, 2013 IBM Corporation

# Companion Piece – z/VM Performance Metrics

- Lists the top 50+ metrics that we find useful, along with descriptions on them
- Where appropriate, includes which:
  - monitor record contains the information
  - Performance Toolkit report displays the information
  - OMEGAMON XE workspace for managing the information
- Rules of Thumb given in some cases

- **Total Processor Utilization:** (Monitor: D0/R2; Toolkit: FCX100 CPU; OMEGAMON: System workspace under headings of Percent CPU). This is the processor utilization from the VM perspective and includes CP, VM System, and Virtual CPU time. It is often beneficial to break this down into the three components:
  - **System Time:** This is the processor time used by the VM control program for system functions that are not directly related to anyone virtual machine. This should be less than 10% of the total processor time for the z/VM LPAR.
  - **CP Processor Time:** This is the processor time used by the VM control program in support of individual virtual machines.
  - **Virtual Processor Time: (Emulation Time):** This is processor time consumed by the virtual machine and the applications within it.

# More to Performance

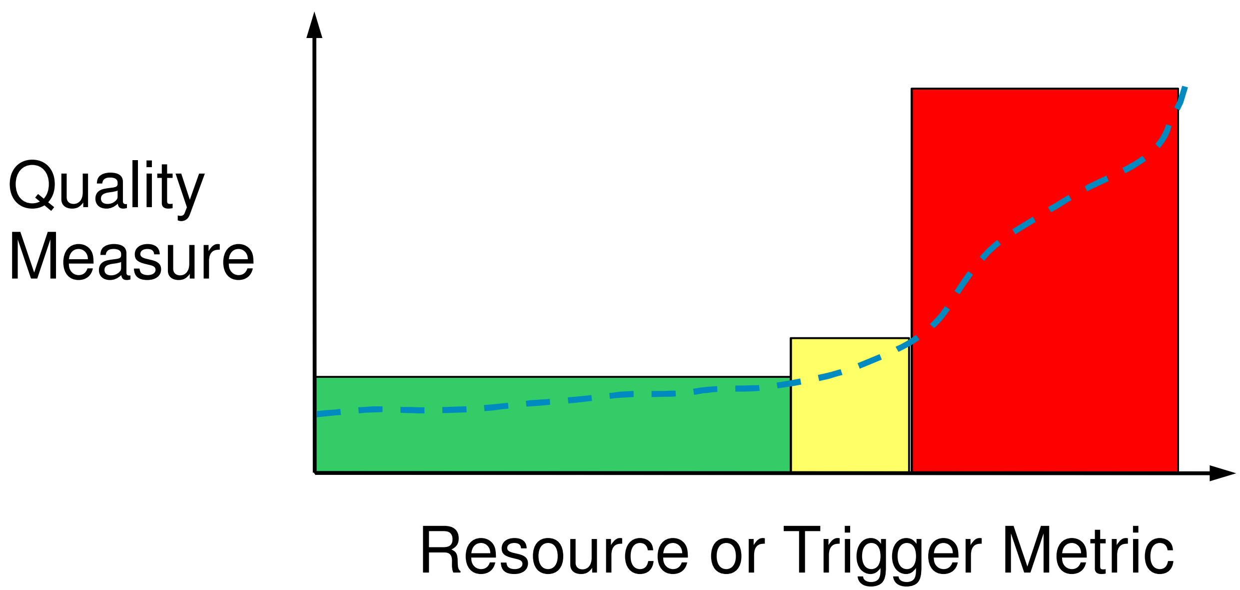


# Real vs. Virtual Planning

- Capacity planning for 'real' resources is one thing, but how do we incorporate 'virtual' resources?
- Need to address “Overhead” or management costs
  - Define 'overhead'
  - z/VM Control Program processor time?
  - System Management virtual machines?
- Virtual ≠ Free
- Need to address peaks
  - Averages alone are not sufficient
- Define what is acceptable – “capacity lag” or impact to SLA
  - Acceptable overcommitment of resources is very dependent on:
    - Workload
    - Environment
    - SLA

# Looking at Resources

- Utilization Metrics: metrics of interest to determine utilization & distribution of resources
- Indicator Metrics: metrics that relate to thresholds or the degree of constraint and pain the system is expressing
- Quality Measures: something that indicates workload and response time and whatever is important to the business
  - Well defined and defined throughout all the disciplines
  - Something that can be mapped to other metrics to indicate a sweet spot



# Real Processor Resources

- Real Processor Resources are perhaps easiest to measure and manage
- Utilization Metrics:
  - LPAR Overhead Time
  - System CPU Time
  - CP CPU time associated with virtual machines
  - Virtual CPU time associated with virtual machines
- Indicator Metrics:
  - System Spin Time (Wall clock, not processor measure)
  - LPAR Suspend Time
  - CPU Wait
- Need to handle Specialty Engines
  - Measure each type
  - Mixed speeds?
  - Changing speeds?
- Keep in mind processor resource limits can pop up elsewhere
  - e.g. Both ends of a HiperSockets connection require processor
- Compare or prorate based on workload (transaction rates).

# Virtual Processor Resources

- Utilization Metrics:
  - CP CPU Time
  - Virtual CPU Time
  - Total CPU Time
- Potentially also include:
  - Processes within Linux
  - Linux Steal time
  - At very least have the above available from Performance Engineering for comparison if z/VM totals look abnormal.
- Again, make accommodations for Specialty Engines
  - Real and Virtual
- Indicator Metrics:
  - CPU Wait
  - Diagnose x'44'
  - Diagnose x'9C'

# Linux View: %Steal

- Current Linux distributions (RHEL 5 & SLES 10) report %Steal as well as pct User and pct System
- %Steal: Linux view of percent of time that it had work to run but was unable to run.
  - z/VM was dispatching other virtual machines, compare to %CPU Wait in z/VM state sampling.
  - z/VM was executing on behalf of the Linux virtual processor, compare to CP CPU usage of the Linux virtual machine
  - Linux yielded its time slice to z/VM via diagnose 0x9C instead of spinning on a formal spin lock. Examine diagnose rates.
  - The z/VM partition was unable to run due to another logical partition being dispatched at the LPAR level.

# Other Processor Planning Thoughts

- Need to have some measure of work or throughput
- Best to determine cost per <something meaningful to everyone>
  - Performance Engineering
  - Business Planning
  - Performance Management
  - Capacity Planning
- Establish in Performance Engineering testing what the target cost / transaction
- Bring in Business Planning to determine target or range of transaction load
- Capacity Planning projects requirements based on above two
- Performance Management folks can help identify problems when things do not track.
- This is a continuous process
- I prefer computing CPU seconds, but if you want to convert to some “MIPS” number or “IFLS” or “Computing Units” feel free. Just make sure everyone uses the same conversion numbers.

# Real Memory

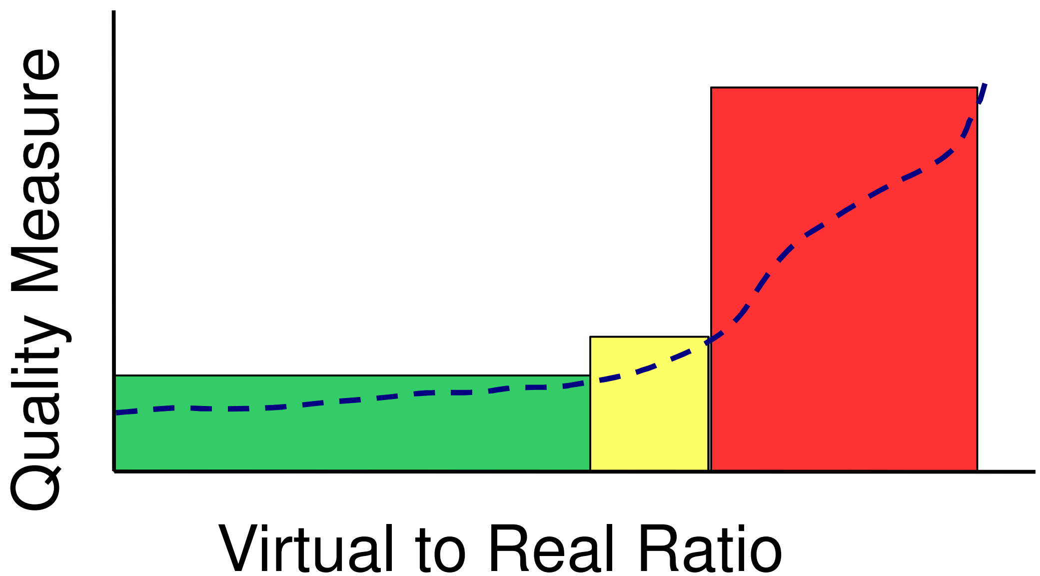
- Utilization Metrics:
  - NonPageable
  - Pageable
  - Minidisk Cache
  - Misc
- Don't forget Expanded Storage for systems prior to z/VM 6.3
- Indicator Metrics:
  - Prior to z/VM 6.3:
    - Emergency Scan on Demand Scan
    - Emergency Scan failures
  - Available List(s) going empty

# Virtual Memory

- Types of Virtual Memory:
  - Virtual Machines
  - NSS/DCSS
  - Virtual Disks in Storage
  - PTRM and other System Utility Spaces
- Virtual Machine Utilization Metrics:
  - Defined virtual memory
  - Backed virtual memory
  - Resident virtual memory
  - Estimated WSS
- Indicator Metrics:
  - Paging Rates (Reads and Writes)
  - Loading User
  - Page Wait (Asynchronous and Synchronous)
- A guest pages may exist on both DASD and Real Memory
- Private DCSSs are considered part of the virtual machine for most metrics, while Shared DCSSs have their own metrics.

# Memory Overcommitment

- Gather data to determine a curve such as below
  - Performance Engineering
  - Tracking Production
  - Artificially limiting amount of real memory
- Result is a Virtual to Real ratio for your workload that is edge of green/yellow.
  - For example, lets say it is 1.8. If you are going to increase workload by adding 30GB of virtual, then you need to add real memory to keep the ratio at 1.8 or lower.



# Other Thoughts on Memory Planning

- What is the right 'over commitment' number? It depends.
  - Definition of virtual and real in the question.
  - Constraints of SLAs
    - All transactions sub-second vs. 99% of transactions sub-second
  - How much can performance features improve things
    - Workloads that are sized poorly at start have more room for improvement
  - Paging configuration
- See http://www.ibm.com/vm/perf/tips/memory.html for additional information

# Network

- Session of its own
- Real Level
- Virtual Level
- Link Aggregation
  - Aggregates total sessions across multiple OSD chpids
  - Does not spread load of a single TCP/IP application session across those chpids.
- Beware of measuring a single session as multiple sessions is where the bandwidth value is.
- Limits/Thresholds/Quality Measures
  - Buffer Overflow counters

# Real I/O

- Utilization Metrics:
  - Channel Utilization
  - Device I/O Rates
    - System I/Os
    - User Driven I/Os
  - Device Utilization
  - Access Density (I/Os per GB space)
- Indicator Metrics:
  - Device Queuing
  - Error Rates
  - IOP Statistics

# Virtual I/O

- Utilization Metrics
  - Virtual I/O per Guest
  - I/Os avoided due to MDC or VDisk
- Indicator Metrics
  - Various levels in software stack where I/O queuing can occur
  - Virtual I/O to Virtual CPU Ratio
- Caution:
  - %IOA (Asynchronous I/O Wait) in Performance Toolkit and similar field in Linux includes time of I/O processing.
    - I/O is relatively slow
    - %IOA may only show up when there is CPU activity or wait on CPU, so a high %IOA isn’t necessarily bad.

# SSI: Capacity Planning

- Great flexibility in managing multiple LPARs
  - Previously, if you split work across LPARs and had an imbalance, it was more difficult to rebalance
  - With SSI, virtual machines can run anywhere in the cluster without a lot of additional work
- Greater responsibility in planning, at two levels
  - Individual members
    - Need to ensure sufficient capacity and resources for the workload on each member
    - Track growth in requirements to limits of the member
  - Cluster-wide
    - Track growth in requirements of overall cluster to the limits of that cluster
    - Need to ensure sufficient white space for planned outages where LGR will be used to move workload out of a given member.

The “Getting Started With Linux” book has been updated with SSI and LGR planning tips.

# SSI & LGR: Planning Relocations

- Need white space for planned outages where you move work off of a given member.
- How will work move off the member?
  - Use existing HA solutions to redirect work to existing servers on other members or elsewhere in enterprise.
  - Use LGR to move to another member.
  - Log off and then logon to another member.
  - Shutdown non-critical virtual machine for duration of the planned outage.
- To where do you move the virtual machines?
  - To a single member or multiple members?
  - To a member on same CEC or another CEC?
  - To a member held in reserve (such as a DR LPAR)?
  - It’s not just one z/VM image anymore

# Other Considerations for Planning

- “The bucket gets heavier as you add water.”
  - Destination system may become more constrained as you continue to relocate virtual machines to it.
- “Get the big rocks in first.”
  - In general, it is better to move the virtual machines generating the greatest memory load first.
    - Larger virtual machines
    - Virtual machines with higher page change rate

# SSI & LGR: Planning White Space

- CPU
  - Shared logical processors?
  - Adjust LPAR weight settings?
  - Vary on additional engines?
- I/O
  - Ensure sufficient resources at all levels:
    - Channel, switch, control unit, device
  - Shared channels?
- Memory white space is not as easy to manage
  - Ensure sufficient paging space and concurrency or data rate capability
  - Increase real memory over commitment?
  - Temporarily decrease size of some virtual machines?
  - Use Dynamic Memory Upgrade?
    - No downgrade available

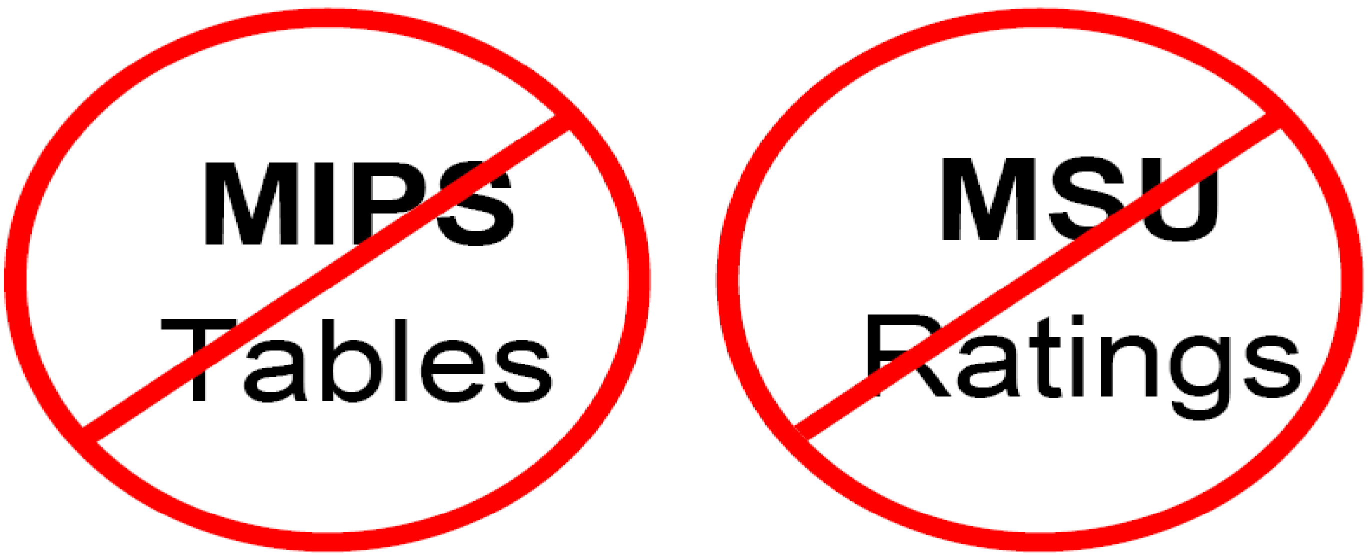
# Data Collection Considerations

- Keep all groups in mind and in agreement
  - The value of your data increases when it can be combined with other data.
- Volume of data
- Retention time
- Granularity or interval of data
- Correlation with other data
- Time zone considerations
- Terminology

# Methods of data collection

- Do It Yourself
- Performance Toolkit Summary/Trend/Histlog
- IBM Tivoli OMEGAMON XE on z/VM and Linux
- Shipping to z/OS

# z10 Capacity Planning in a nutshell

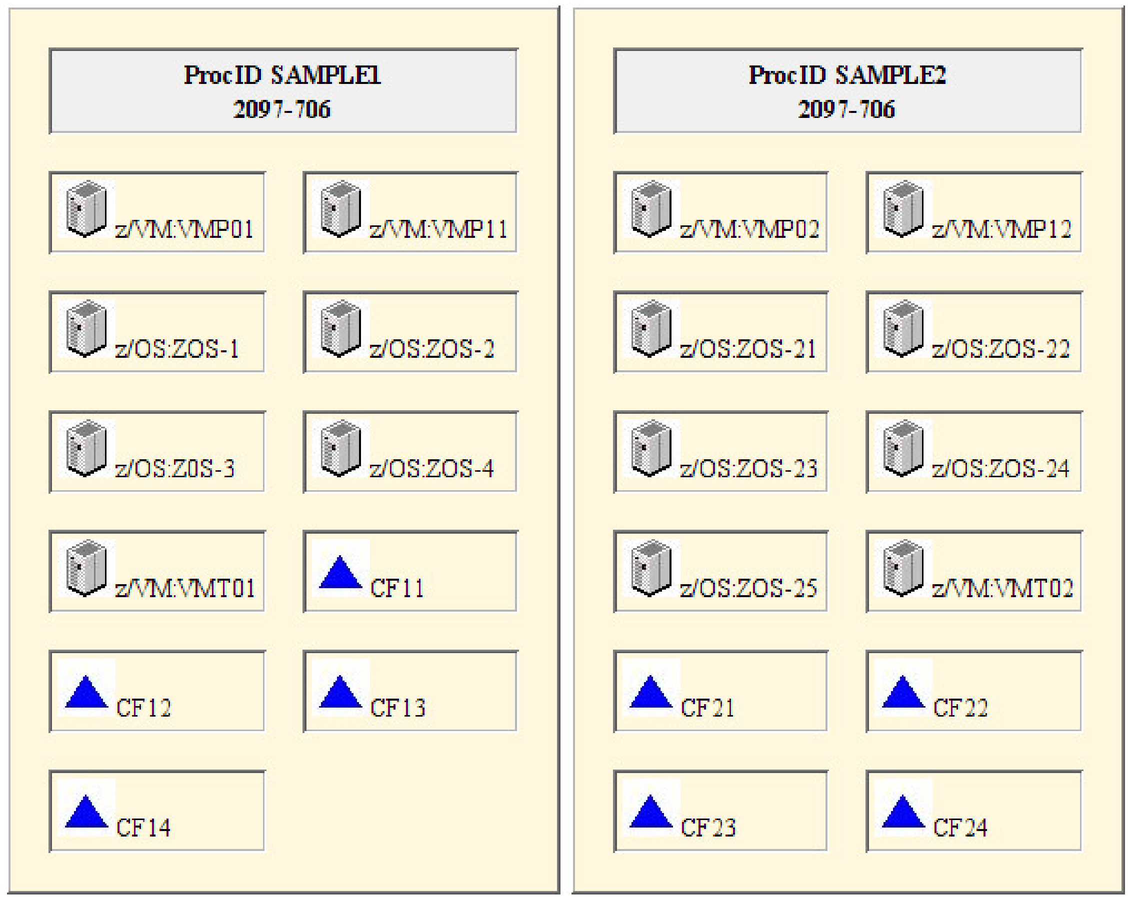



**Don’t use “single-number tables” for capacity comparisons!**

**Use zPCR and/or zCP3000 to model before and after configurations**
**Work with IBM technical support for capacity planning!**
**Customers can now use zPCR**

# IBM Techline: Complete topology from z/VM data

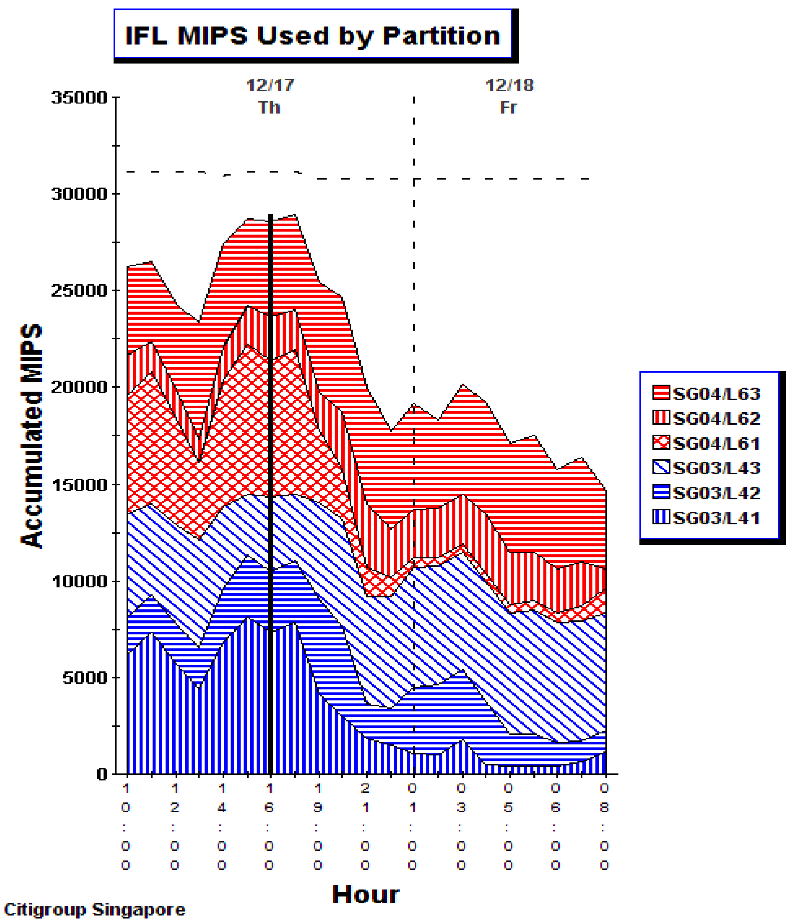
## Processor View for Generic Customer



# IBM Techline: Example of CPU Analysis

Selected Partition IFL MIPS by Sample : zCP3000 Graph and Text: INV1011

File Action




This graph shows the accumulated IFL MIPS consumed for each selected ProcID/Partition by sample.

| | Avg. | Partition | | Accumulated | | Study Interval MIPS | |
|---|---|---|---|---|---|---|---|
| ProcID/Partition | IFL | Avg. | Max. | Avg. | Max. | Value | Accum. |
| SG04/L63 | 11.0 | 5,207 | 6,098 | 5,207 | 6,098 | 4,908 | 4,908 |
| SG04/L62 | 6.9 | 2,222 | 3,266 | 7,429 | 9,363 | 2,305 | 7,213 |
| SG04/L61 | 16.0 | 3,116 | 7,787 | 10,544 | 14,426 | 7,033 | 14,246 |
| SG03/L43 | 11.0 | 5,375 | 6,433 | 15,919 | 18,109 | 3,864 | 18,109 |
| SG03/L42 | 7.0 | 2,581 | 4,928 | 18,500 | 21,697 | 3,112 | 21,221 |
| SG03/L41 | 16.0 | 3,422 | 8,147 | 21,922 | 28,909 | 7,355 | 28,576 |

28,909 is the maximum number of MIPS being used today.

The accumulated columns represent the average and maximum for the ProcIDs/Partitions listed so far. So the second line would be the average and maximum for the first two entries. Hence the bottom line, would be the average and maximum for all partitions together.

The study interval is 12/17/09 at 16:00 and is the 90th percentile interval from the Prime shift. This is shown on the graph as a solid line. The value at this point is 28,576 MIPS.

The MIPS ratio of the maximum to the study interval is 1.01. This means that when the study interval grows to the maximum value, the peak would be about 29246. Keep in mind that this is very fuzzy. You should look at this as a value ±5% or on the interval [27784, 30708] MIPS.

The reference processor is set to a 2094-701 with an assumed capacity of 602.0 MIPS.

# IBM Techline: Memory Summary Example

File | Analysis

System | Workloads | Memory | Performance

Express Memory Sizes As: ◉ MB ○ Pages

| Description | Virtual Memory | CMMA Active | WSS Intv | WSS Min | WSS Max | Memory Used Intv | Memory Used Min | Memory Used Max |
|---|---|---|---|---|---|---|---|---|
| LXPS3061 | 1,707 | 0 | 1,491 | 409 | 1,707 | 1,259 | 346 | 1,504 |
| LXPS3095 | 4,608 | 0 | 4,096 | 3,772 | 4,291 | 3,406 | 3,145 | 3,581 |
| LXPS3093 | 11,750 | 1 | 11,428 | 11,423 | 11,750 | 9,533 | 9,522 | 9,877 |
| LXPS3033 | 4,736 | 0 | 4,736 | 4,341 | 4,736 | 4,662 | 3,031 | 4,708 |
| LXPS3139 | 5,120 | 1 | 5,120 | 5,120 | 5,120 | 4,605 | 4,443 | 4,891 |

| | Interval | Min | Max |
|---|---|---|---|
| Virtual Memory Sum | 53,521 | 53,521 | 53,521 |
| WSS Total | 52,472 | 44,576 | 52,862 |
| DPA | 142,090 | 142,082 | 142,090 |
| Memory Utilization % | 36.9% | 31.4% | 37.2% |
| Memory Overcommit | 0.4 | 0.4 | 0.4 |
| Available Queue | 9 | 6 | 39,017 |
| CS<->ES Page Rate | 603 | 2 | 603 |

**LPAR ES 20,480 MB**
**LPAR CS 143,360 MB**

Cancel | Apply

# IBM Techline Support

- **IBM Techline Support – z/VM Capacity Planning**
  - **http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS2875**

- **Contact your IBMer for in-depth analysis**

- **See free tools such as zPCR for processor sizing**
  - **http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/PRS1381**

- **Thanks to following for info on Techline and ATS Offerings:**
  - **Gretchen Frye**
  - **Liz Holland**

# Summary

- You have to Plan to do Capacity Planning if you want to do it successfully
- Otherwise, it becomes Capacity Scrambling
- Lots of resources available to help from IBM and Others

Contact Information:

Bill Bitner

IBM Endicott

bitnerb@us.ibm.com
+1.607.429.3286