z/VM Guest Performance

WAVV 2004

Bill Bitner
IBM Endicott
bitnerb@us.ibm.com

Last updated: April 24, 2004

Legal Stuff

The information contained in this document has not been submitted to any formal IBM test and is distributed on an "as is" basis without any warranty either express or implied. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environment do so at their own risk.

In this document, any references made to an IBM licensed program are not intended to state or imply that only IBM's licensed program may be used; any functionally equivalent program may be used instead.

Any performance data contained in this document was determined in a controlled environment and, therefore, the results which may be obtained in other operating environments may vary significantly.

Users of this document should verify the applicable data for their specific environments.

It is possible that this material may contain references to, or information about, IBM products (machines and programs), programming, or services that are not announced in your country or not yet announced by IBM. Such references or information should not be construed to mean that IBM intends to announce such IBM products, programming, or services.

Should the speaker start getting too silly, IBM will deny any knowledge of his association with the corporation.

The following are **Trademarks** of the IBM Corporation:

VM/ESA, e-business logo*, HiperSockets, IBM*, IBM logo*,
IBM eServer, RAMAC*, TotalStorage, z/OS, z/VM, zSeries

LINUX is a registered trademark of Linus Torvalds

Overview

- General management of resources
- Processor
- I/O
- Storage and Paging
- Linux® guidelines
- Performance Monitoring

What do you mean by "Performance?"

- ITR = Internal Throughput Rate = a measure of work per CPU second
- ETR = External Throughput Rate = a measure of work per wallclock second
- CPU Utilization = how busy processor is; tied to ITR
- Response Time (Elapsed Time) = how long jobs take; tied to ETR
- Interactive Users vs. Batch Work
- How many phone calls you get

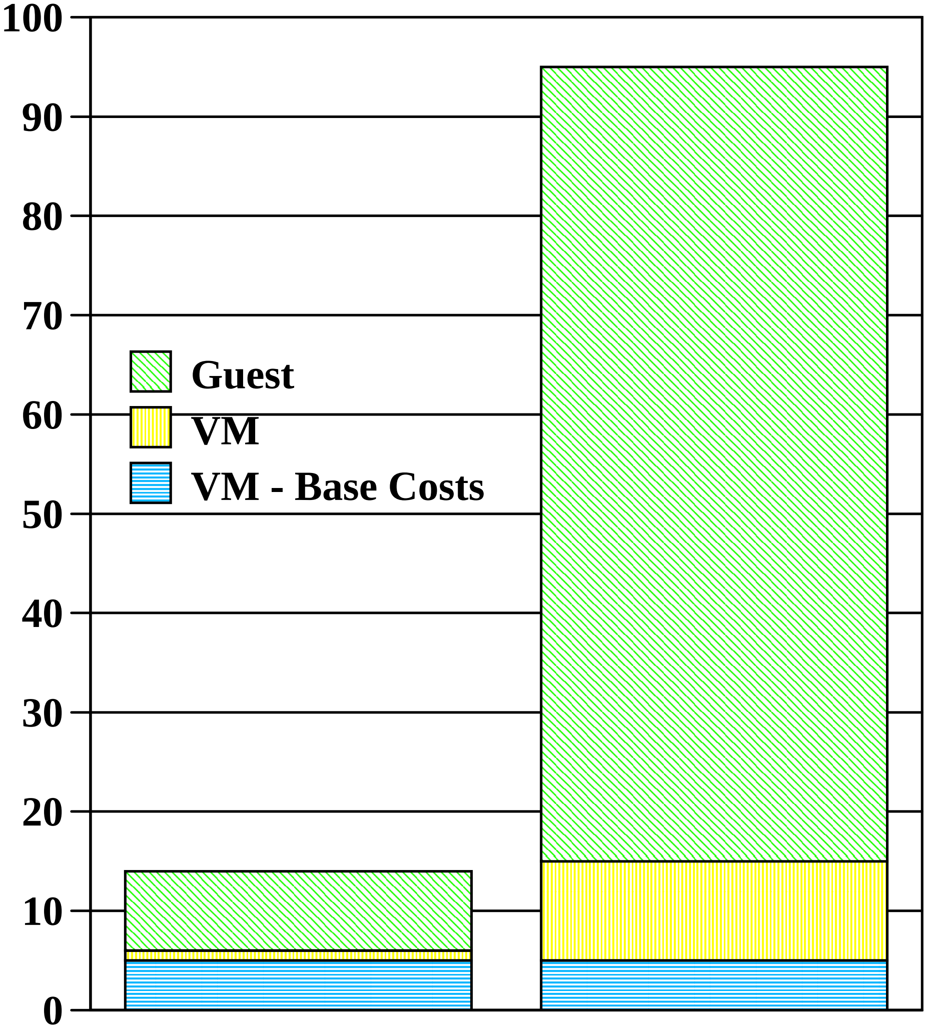
Processor Resources

- Configuration
  - Virtual 1- to 64-way, defined in user directory or via CP command
  - A real processor can be dedicated to a virtual machine
  - Do not recommend use of more virtual processors than there are real
  - Do not recommend mixing shared and dedicated processors
- Control and Limits
  - "Share" setting
  - Absolute or Relative
  - Target minimum and maximum values
  - Maximum values (limit shares) either hard or soft
  - "Share" for virtual machine, divided amongst its virtual processors

Processor Usage by VM

- Base costs and background work
  - Scheduling and dispatching
  - Accounting
  - Monitor
- Costs proportional to guest requests or requirements of VM

Guest Example



Processor: SIE Exits

- SIE = Start Interpretive Execution
- Used by z/VM™ to run a guest
- Exits from SIE indicate work for VM
- Rate of SIE executions available from most performance monitor products (e.g. VMPRF, RTM, etc.)
- Hardware assists can help avoid SIE exits
- Most common reasons for exiting SIE
  - I/O processing
  - Page fault resolution
  - Instruction simulation
  - Minor time slice expires
  - Loaded wait state

Avoiding Exits from SIE

- Data in memory techniques avoid I/O
- I/O Assist avoids SIE exit to handle:
  - I/O interrupt processing
  - CCW translation from virtual to real addresses
- CCW translation bypass for V=R guest
- Minor time slice: SET SRM DSPSLICE
- Avoid Paging
  - V=R/F
  - Reserved pages for V=V
  - Sufficient storage

I/O Resources

- Configuration
  - Dedicated devices (Tape Drives, DASD, Network devices)
  - Partitioned devices (minidisks)
  - Virtualized devices (minidisks, crypto)
  - Simulated devices (Guest LAN, virtual disks in storage)
  - Define or attach dynamically
- Control and Limits
  - Indirect control through "share" setting
  - Real devices can be throttled at device level
  - Priority can be set for virtual machine
    - CP uses to effect queue placement for DASD devices
    - HW uses to effect priority in channel usage
  - Minidisk Cache fair share limits can be turned off for virtual machine

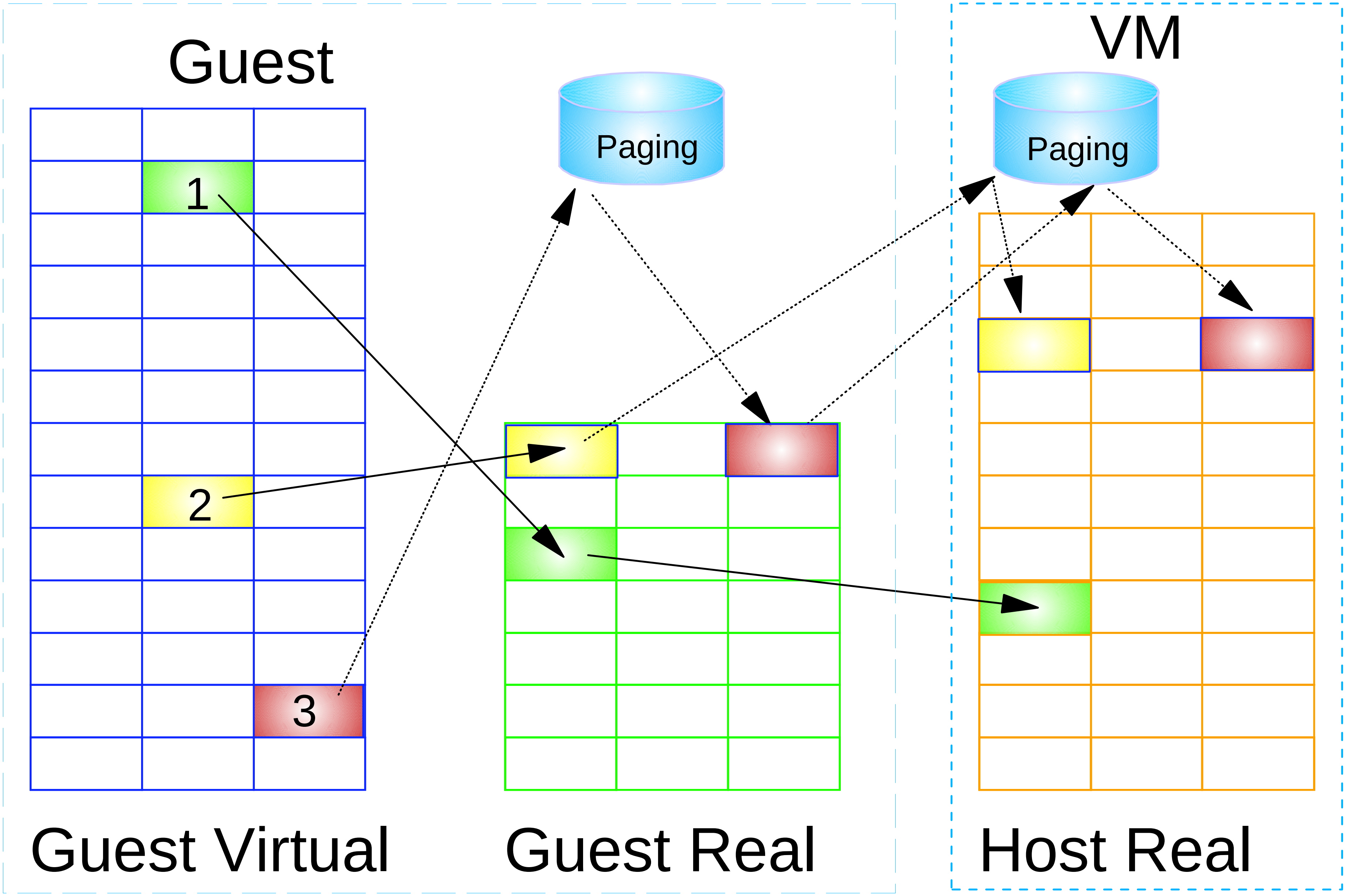
I/O Considerations

- Traditional benefit of V=R/F guests and I/O Assist usually does not apply to Linux guests
  - Integrated Facility for Linux (IFL) processors most often used for Linux
  - IFL requires LPAR which results in loss of I/O Assist
- Dedicated I/O is not eligible for Minidisk Cache (MDC)
- MDC read performance is as good as VM virtual disk in storage performance
- Both VM vdisks and MDC require sufficient storage

Storage Resources

- Configuration
  - Defined in user directory or via CP command
  - Can define storage with gaps (useful for testing)
  - Can attach expanded storage to virtual machine
  - Machine can be V=V, V=F, or V=R
- Control and Limits
  - Scheduler helps control over committing storage and paging resources
  - Virtual Machines that do not "fit" criteria placed in eligible list
  - Virtual Machine can be made exempt from eligible list via QUICKDSP
  - Can "reserve" or "lock" pages for V=V guests
    - Reserve a number of pages to influence storage management page steal algorithms (recommended approach)
    - Lock specific pages (less flexible and forces page below 2GB)

Paging Considerations



Paging Considerations

- For V=V guests the potential exists for "Double Paging"
- No VM paging for V=R/F
- The closer the virtual machine size is to the amount of memory the Linux guest truly needs, the lower the Linux swapping..
  - However, oversizing the virtual machine size for Linux guests has other negative effects
- PAGEX and Asynchronous Page Fault used where appropriate
- VM can use expanded storage for high speed paging device
- There can be an advantage to defining some processor memory as expanded storage
  - See www.vm.**ibm.com**/perf/tips/storconf.html

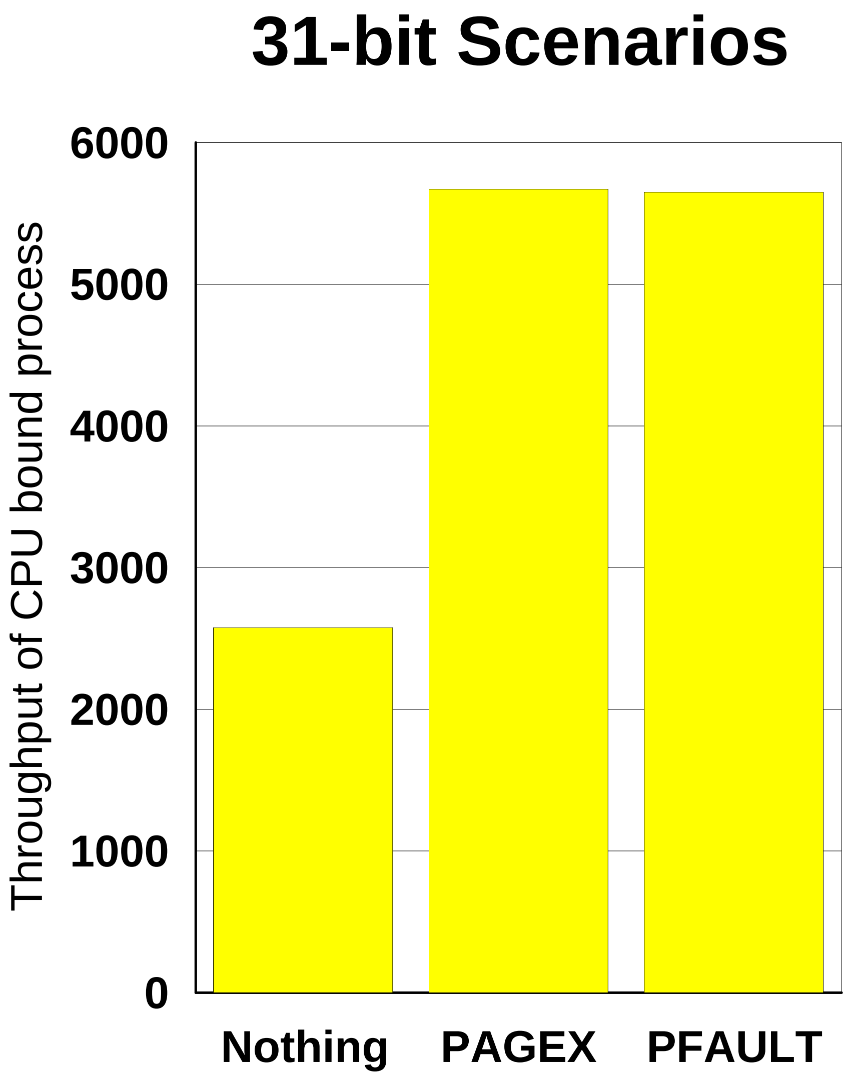
V=R/F/V Considerations

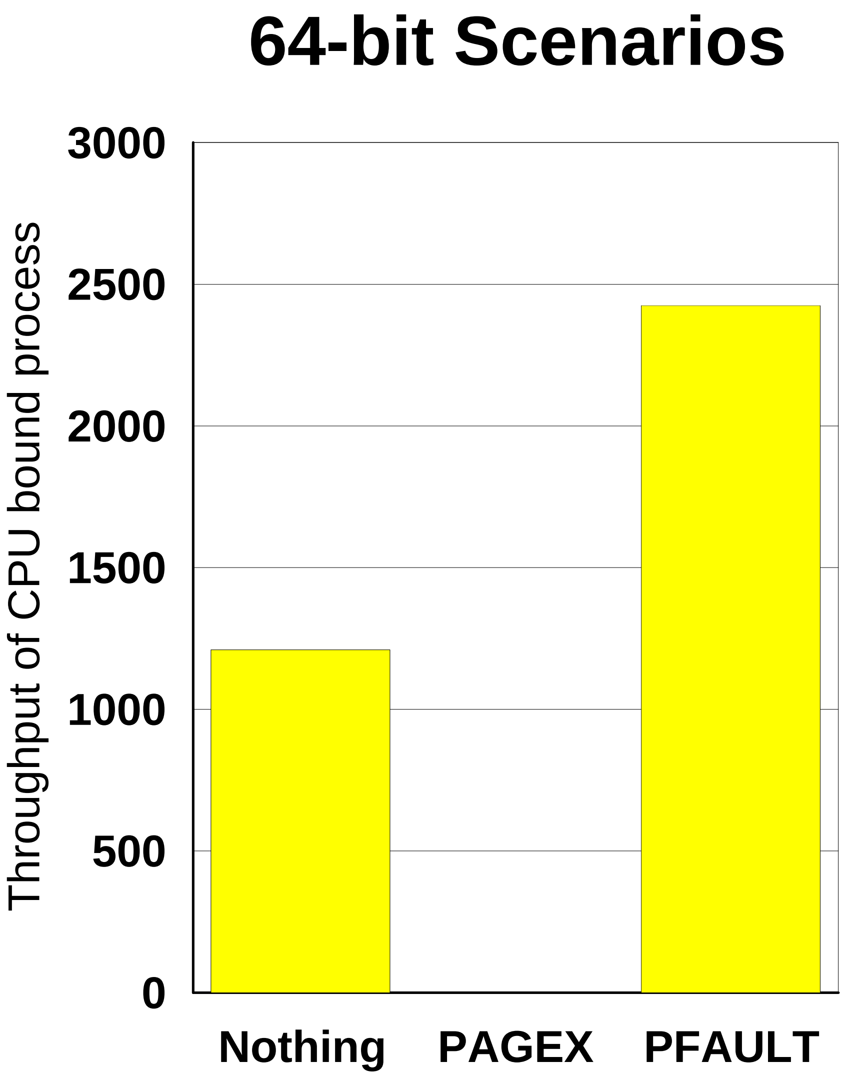
- V=R/F potential I/O assist benefit (saves CPU)
- V=F avoids overhead of recovering V=R
- 1 V=R + 5 V=F or 6 V=F
- V=V avoids dedicating storage
- V=R defaults to dedicating processors
- Running z/VM in an LPAR -
  - No V=F, only V=R, but without I/O Assist
  - Often better to use V=V and reserve pages

Asynchronous Page Fault Facility

- Ordinarily, page faults serialize the virtual machine. This can be a throughput and response time problem for guest systems
- Enhancements designed for Linux
- PFAULT macro
  - Accepts 64-bit inputs
  - Provides 64-bit PSW masks
- Diagnose x'258'
- Older PAGEX interface limited to 31-bit
- z/VM 4.2.0
- Linux 2.4 Kernel required

Page Fault Tests





Virtual MP Support

- Define additional processors dynamically
  - Directory include MACHINE ESA 2
  - CP DEFINE CPU vcpu_addr
- Or put everything in the directory
  - CPU 00 NODEDICATE
  - CPU 01 NODEDICATE
- Detaching vCPU resets virtual machine
- For testing: more virtual than real processors

Virtual MP Support

- CP commands of interest
  - QUERY VIRTUAL CPUS
  - CPU vcpu_addr cmd_line
  - DEDICATE and UNDEDICATE
- Share setting is for virtual machine, divided amongst all virtual processors
- Mixing dedicated and shared processors is not recommended
- Defined but inactive vCPU (stopped state) makes guest ineligible for I/O assist
- Dedicated processor appears 100% busy on various VM performance reports

Linux Guest Guidelines

- Why does my idle Linux consume Processor resources?
  - Timer pops
- Is the number and size of guests important?
  - Yes! It is virtual storage, but it isn't magic. It has to reside somewhere when Linux guest is running.
- How big should my Linux guest be?
  - Not bigger than you need
- Where should Linux swap?
  - Multiple choices: XPRAM, Mdisk, Tdisk, Vdisk
- Should I set QUICKDSP ON for my Linux Guest?
  - Production vs. Test vs. Development machines
- See the following URL for other information: www.vm.**ibm.com**/perf/tips/linuxper.html
- See APAR VM63282 for better dispatch list management

Swapping Configuration

- The trade-off
  - Defining virtual machine too large may cause excess memory to be used inefficiently for file and buffer cache.
  - Defining virtual machine too small may cause swapping which is expensive in processor time and impacts response time.
- Configure so that swap rate is zero or very low.
- Virtual disk in storage can be used to mitigate cost of swapping.
  - Pros:
    - very easy from administration view
    - virtual disk blocks not created unless referenced
    - performance
  - Cons:
    - DAT structures required below 2GB and are not pageable
    - Steal algorithms favor those pages over pages of idle users
    - Disk block pages below 2GB prior to z/VM 4.4.0
- Do not define virtual disks in storage larger than necessary

Networking Choices

- Lots of variations for connecting
  - guests to other guests
  - guests to another LPAR
  - guests to physical network
- Continued improvement in both Linux and VM stacks
- Workload dependent
  - MTU impact
  - Performance may improve as load increases
    - data rate and number of connections

Guest to Guest

- Guest LAN
  - Simulated HiperSockets slightly lower pathlength than Simulated GbE
  - Use with the virtual switch (z/VM 4.4)
- HiperSockets
  - Requires locking real memory below 2GB
  - Configuration limitations
  - Better performance for large data transfers

Guest to Another LPAR

- HiperSockets
  - Best solution
  - Pay attention to MFS (MTU)
- Shared OSA GbE
  - Additional overhead and latency even when shared card

Guests to External Network

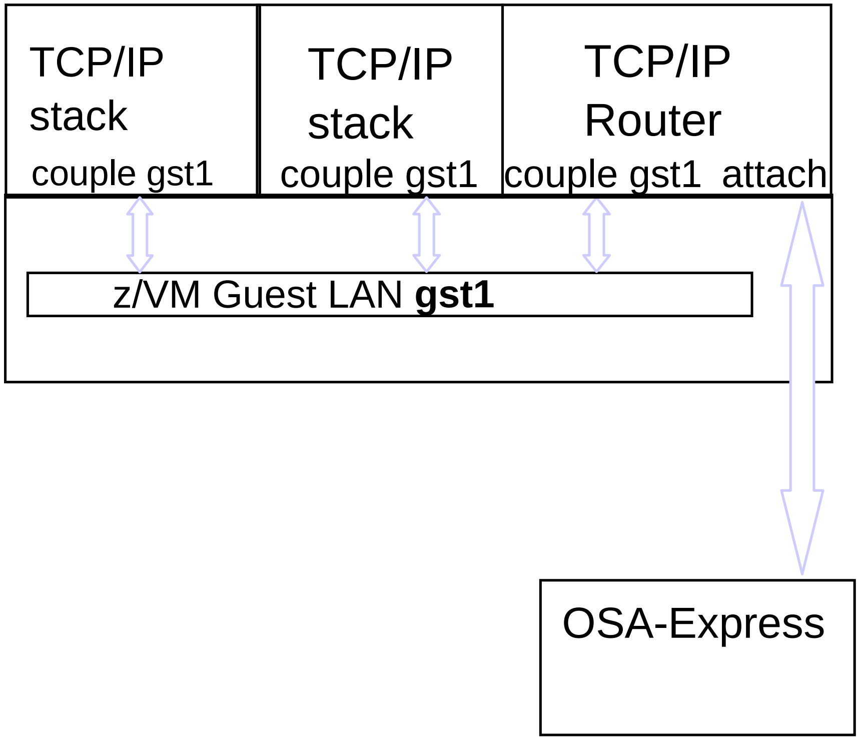
- Guests direct connect to OSA
  - lowest pathlength, especially with z/VM 4.4.0 with hardware that supports Adapter Interrupt Passthru (AIP)
  - requires locking real memory below 2GB
- Virtual switch
  - requires z/VM 4.4.0
- Virtual Machine Router
  - extra pathlength for moving and processing data

Virtual Switch - New in z/VM 4.4.0

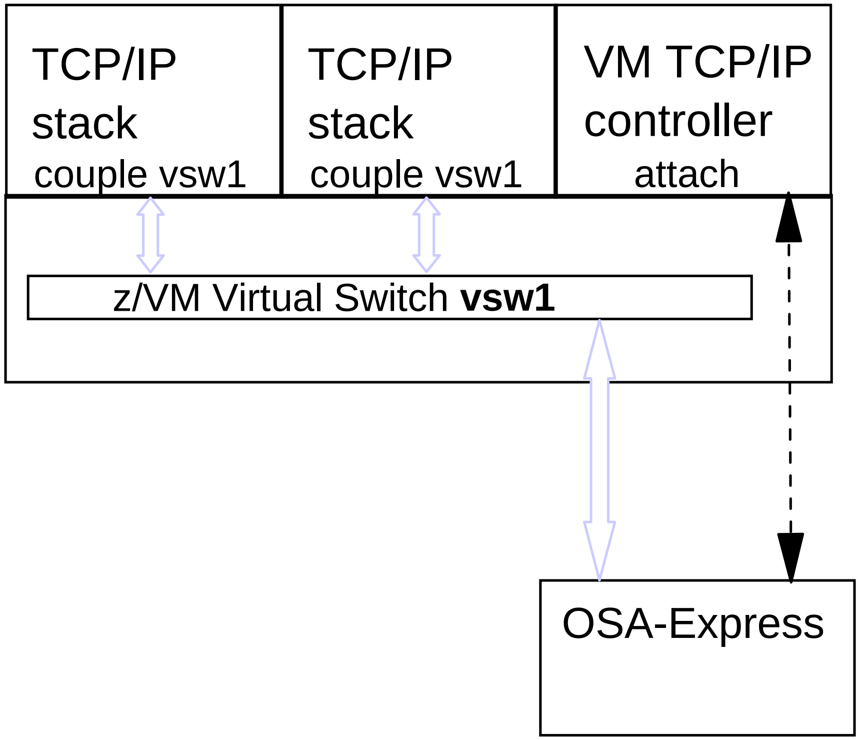
- Layer 3 switch
  - Switches packets between QDIO guest LAN and OSA Express physical network
  - Eliminates need for layer 3 router
  - Supports transparent VLAN specifications for guests connected to Virtual Switch
  - Switching function performed entirely by CP
  - z/VM TCP/IP stack used for setup and control functions
- Provides transparent bridging
  - Learning - automatic configuration of IP addresses
  - Flooding - deliver packets for unknown IP addresses to all stations
  - Aging - forget learned IP addresses after some period of inactivity

Virtual Switch Topology

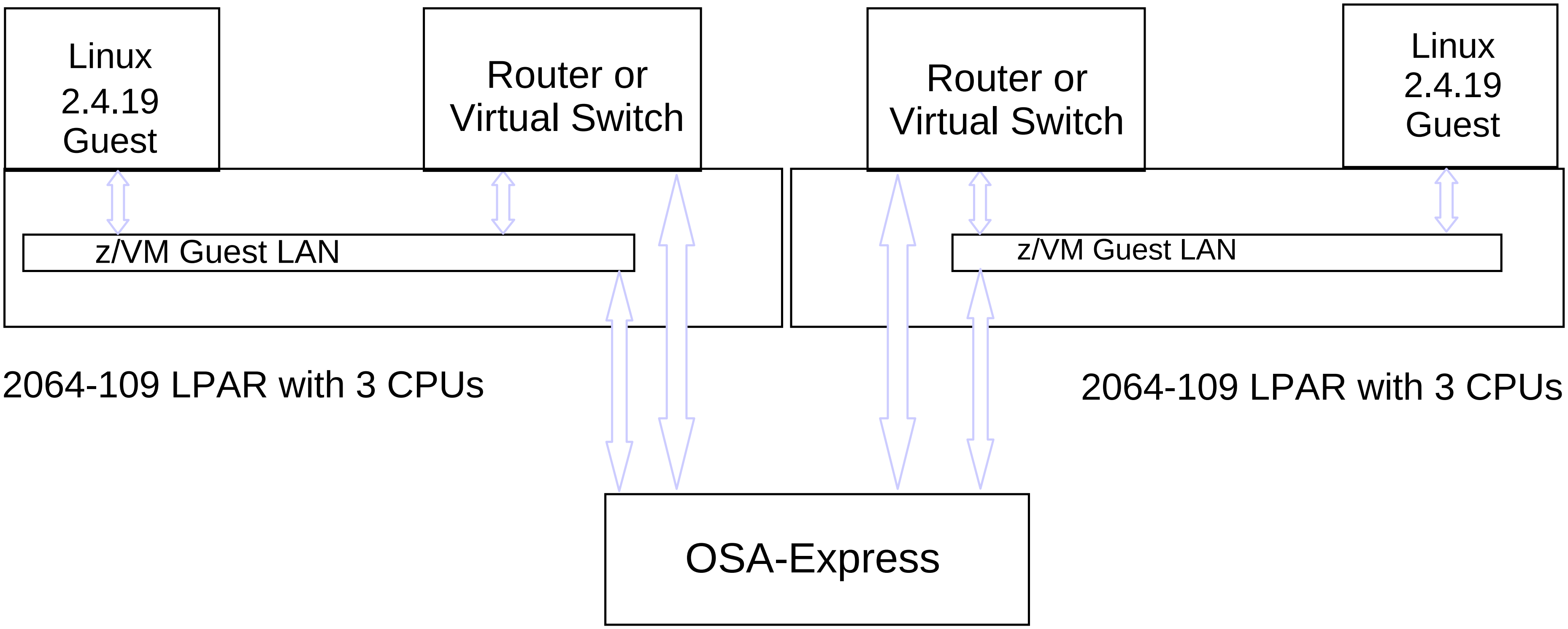
Traditional Guest LAN



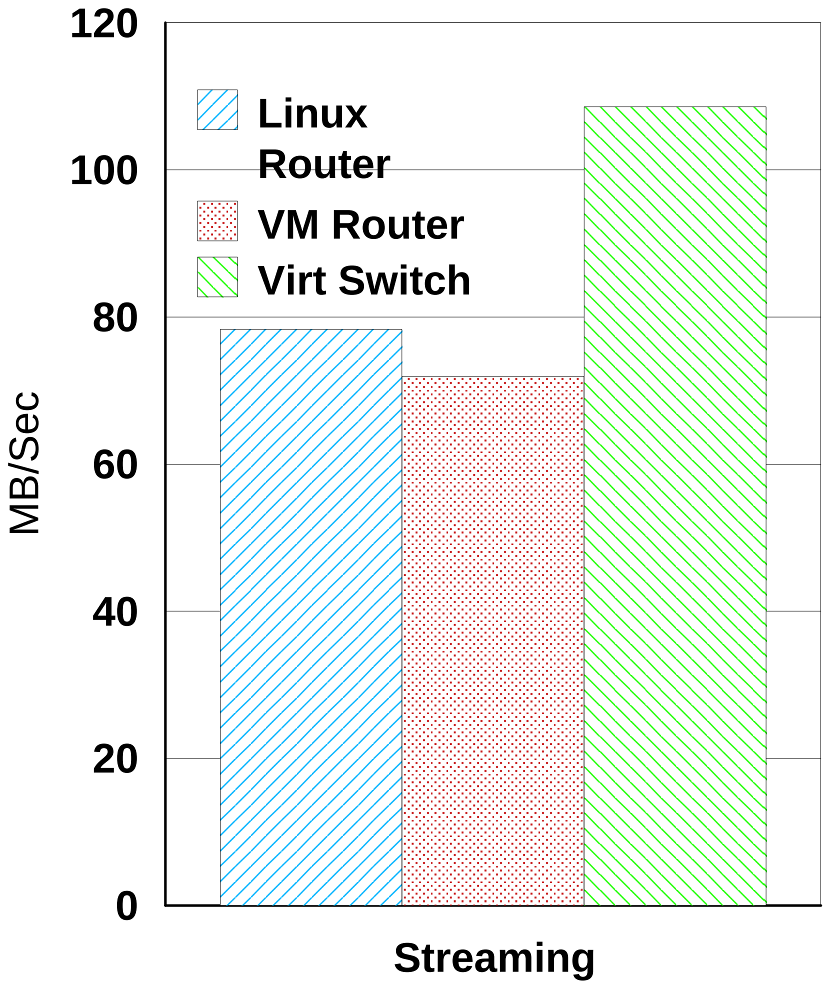
Virtual Switch Guest LAN

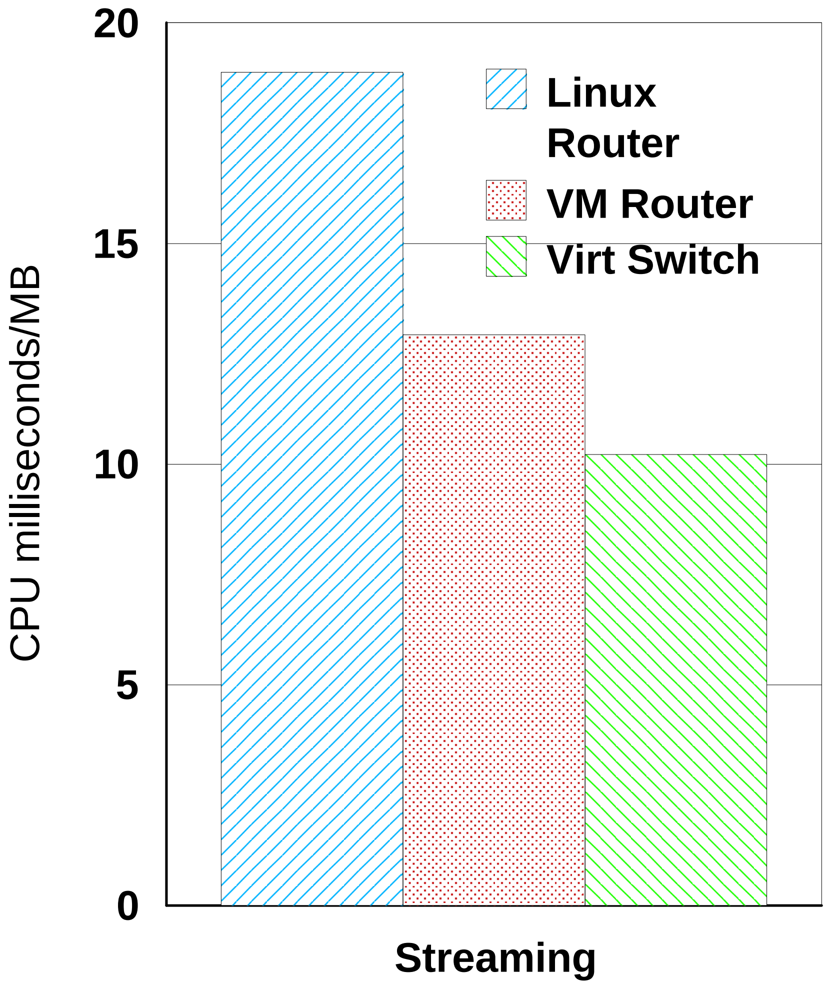


Virtual Switch Test Configuration

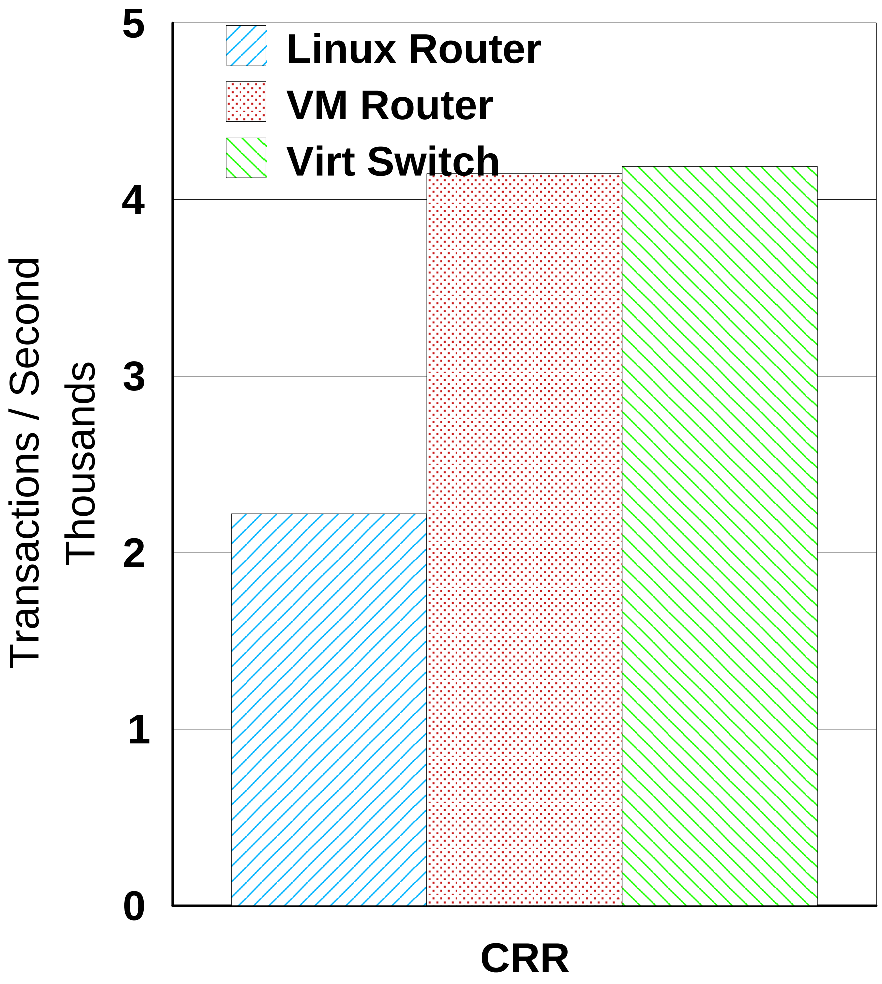


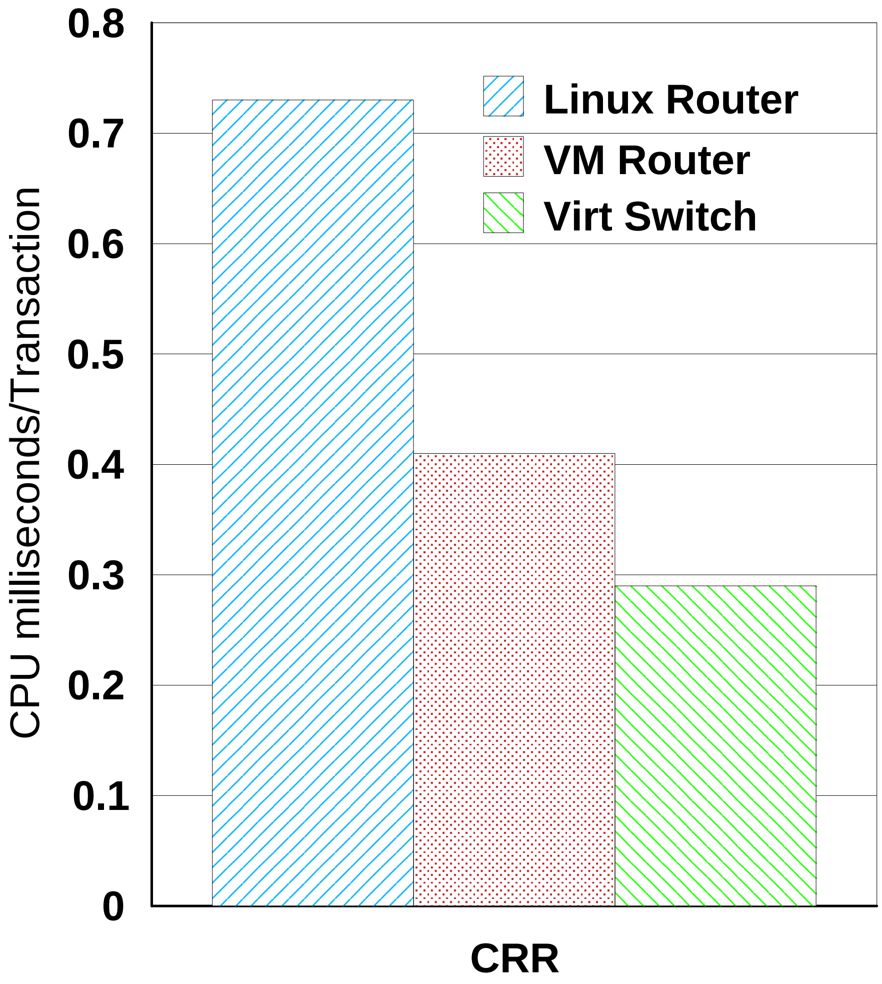
Virtual Switch - Streaming (MTU 8992)



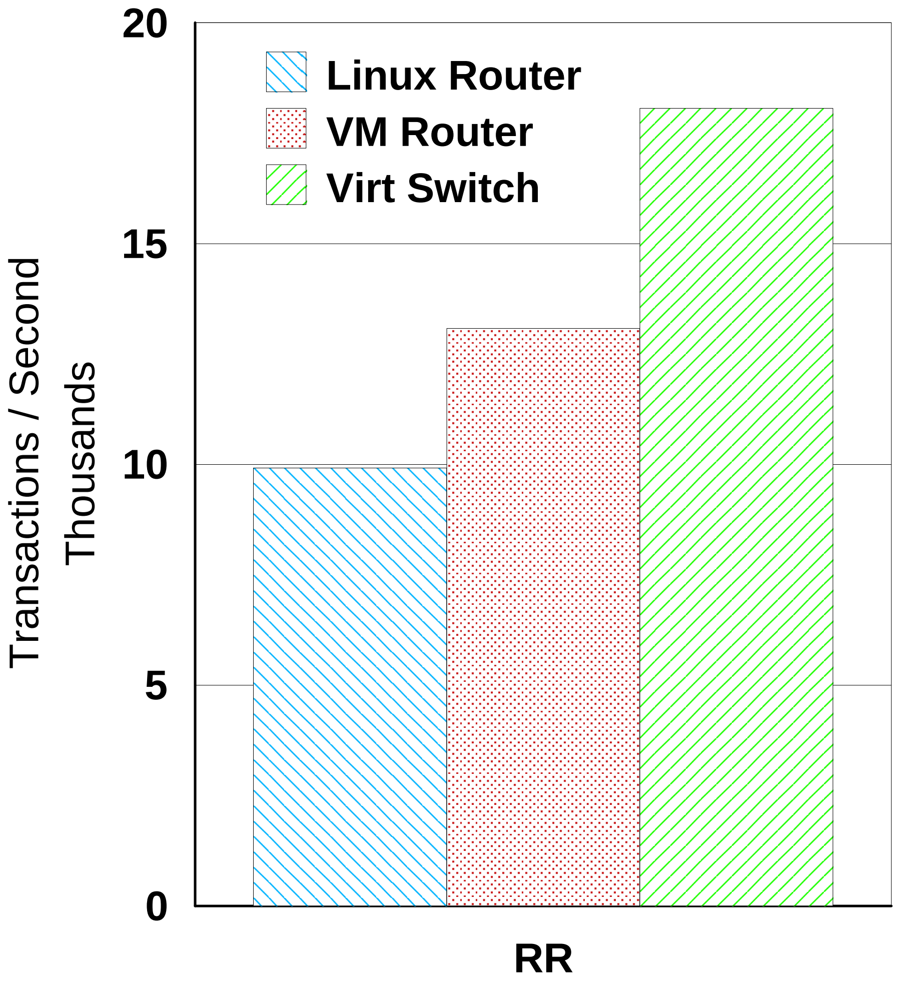


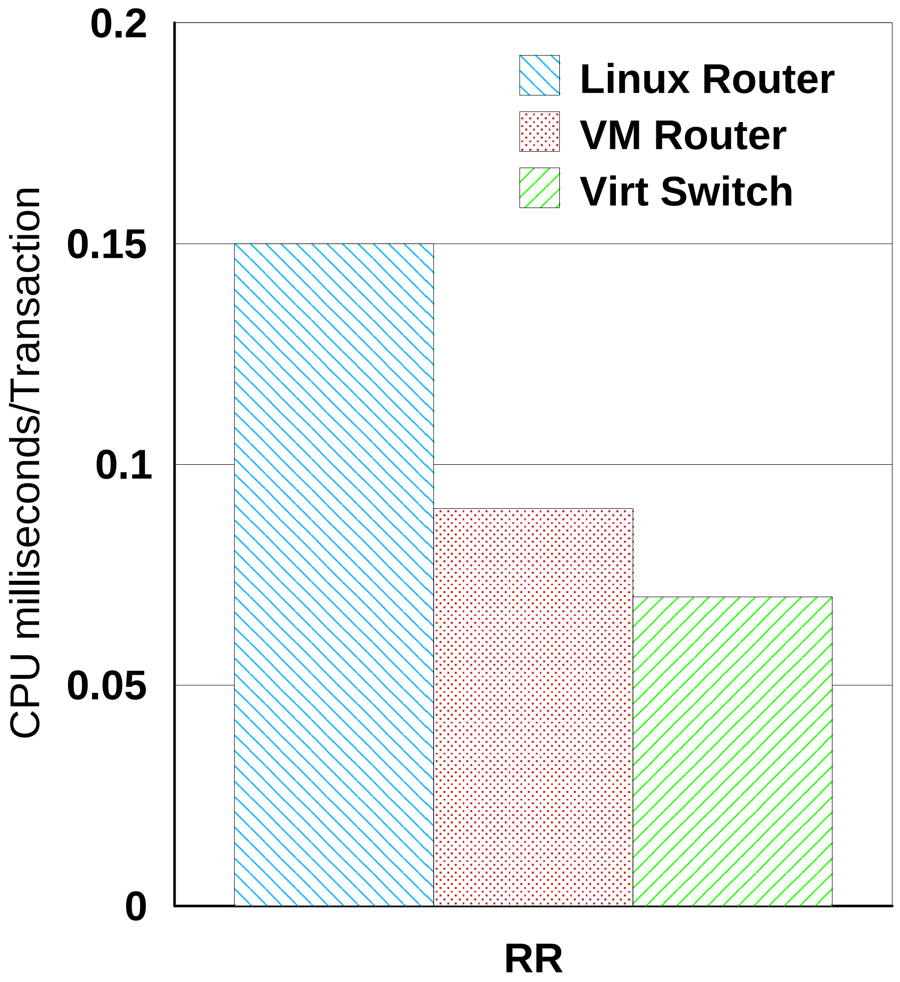
Virtual Switch - CRR (MTU 8992)





Virtual Switch - RR (MTU 1492)



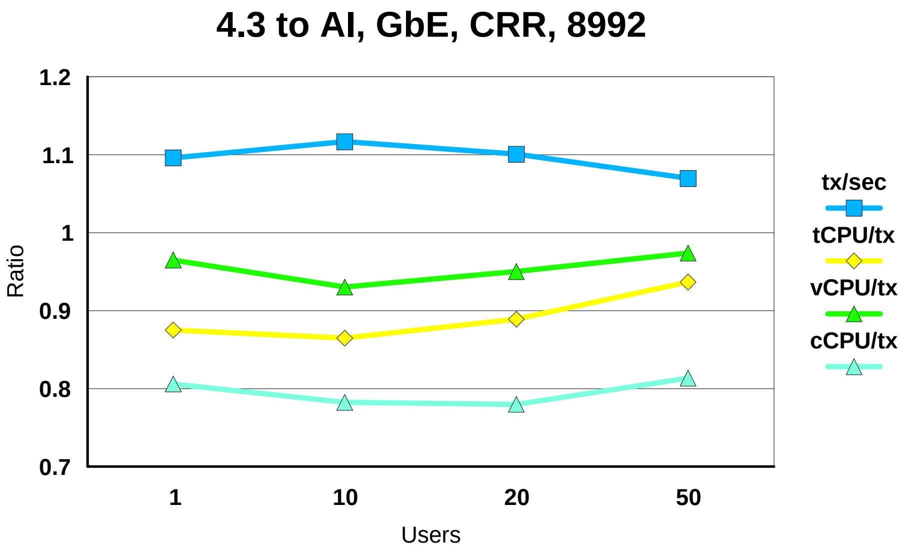


Queued I/O Assist

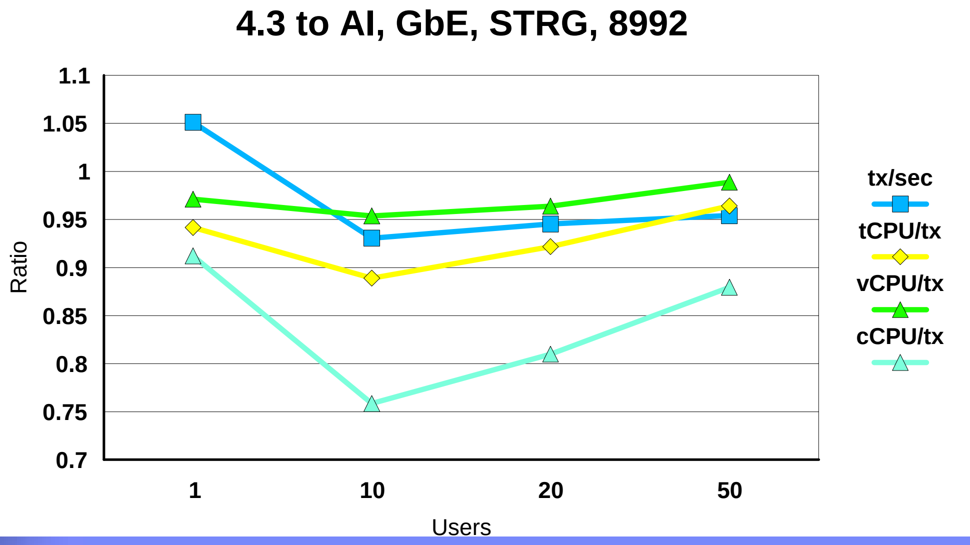
- QDIO devices (FCP, OSA Express, HiperSockets) induce overhead due to high interruption rates
  - z/VM Control Program has to mediate between hardware interruptions and guests
  - As interruption rates go up, this overhead increases
- New hardware facility designed to address this problem
  - Allows interruptions to be presented directly by hardware for active guest
  - Delivers "thin" signal to CP when interruption is for idle guest
  - Extends "thin interrupts" from iQDIO to QDIO and FCP
  - New feature limited to z990 and z890.
- Changes in z/VM and Linux to take advantage of assist
  - QUERY/SET QIOASSIST
- See http://www.vm.ibm.com/perf/aip.html for more information.

PCI to AI, Linux, GbE

Usually it's great!



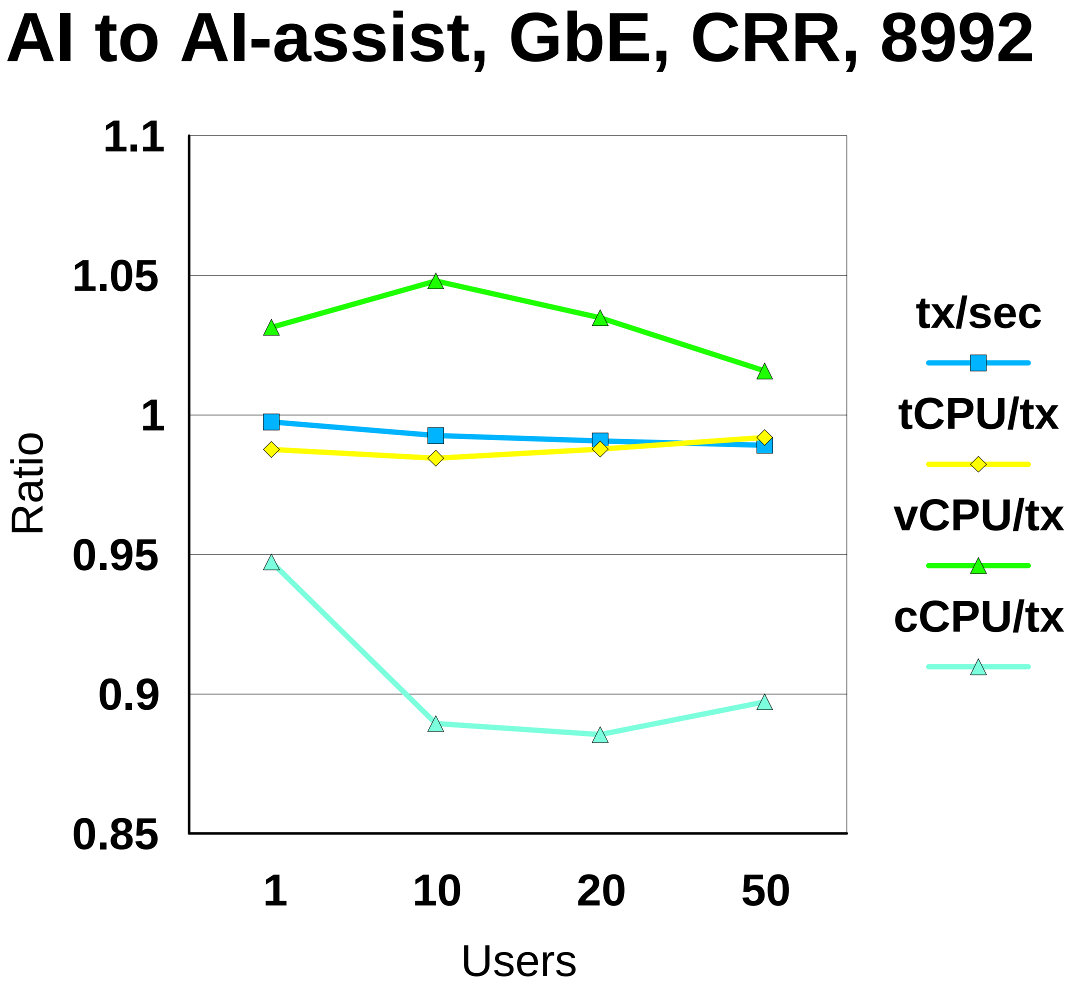
Rarely, it's marginal.



AI to AI-Assist, Linux, GbE

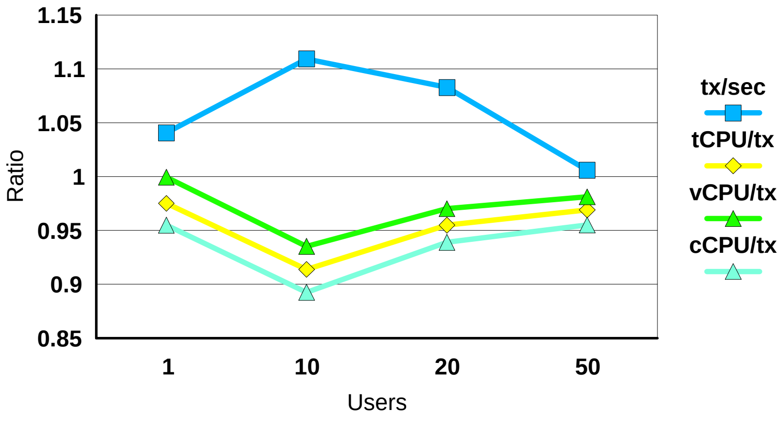
Generally, we see this:

- Tx/sec flat
- Small rise in virtual/tx
- Good drop in CP/tx



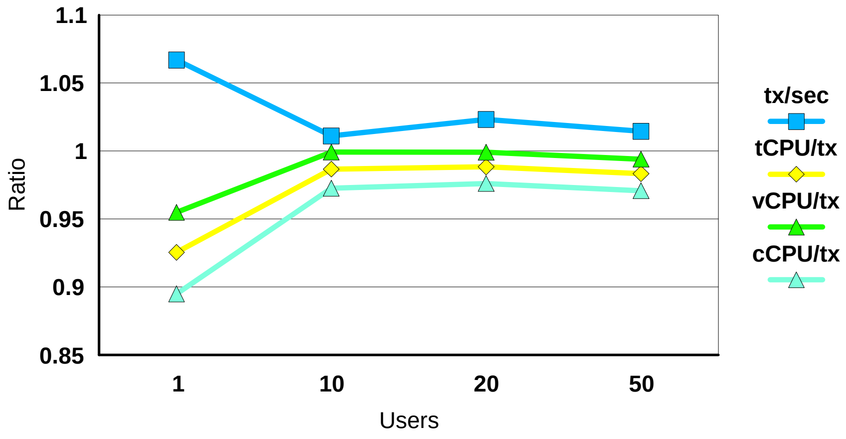
AI to AI-Assist, Linux, HiperSockets

AI to AI-assist, Hiper, RR, 57344



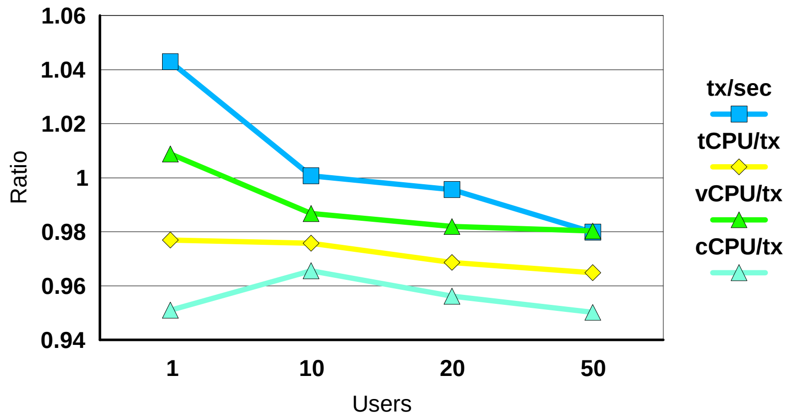
Nice!

AI to AI-assist, Hiper, STRG, 8992



Ho-hum.

AI to AI-assist, Hiper, CRR, 8992

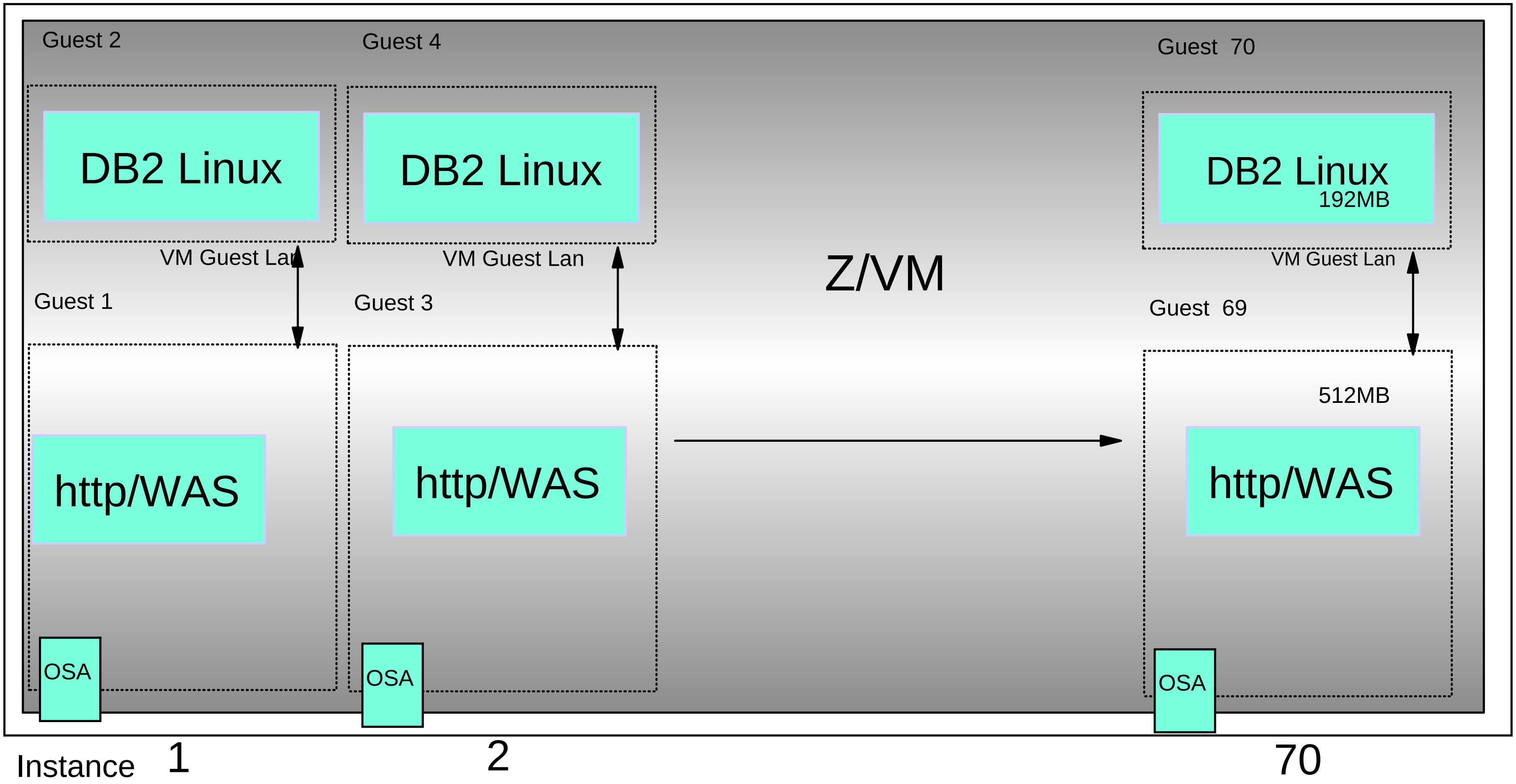


There are only a couple of "oops" cases.

Oops!

Physical Configuration

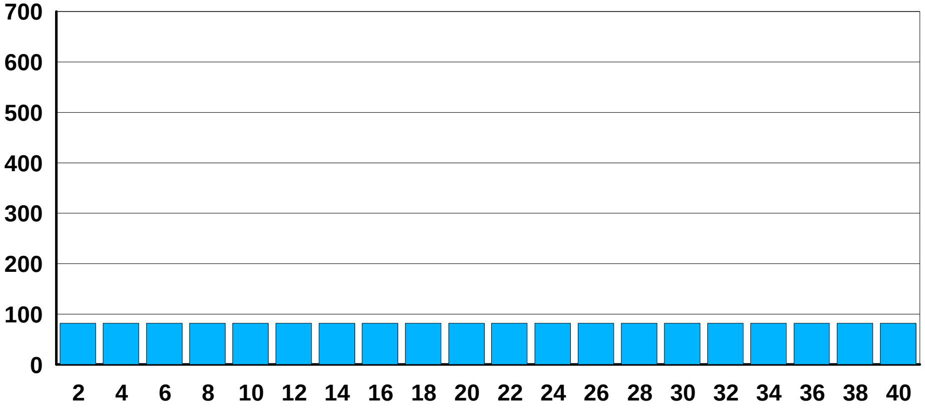
Z900 16 way



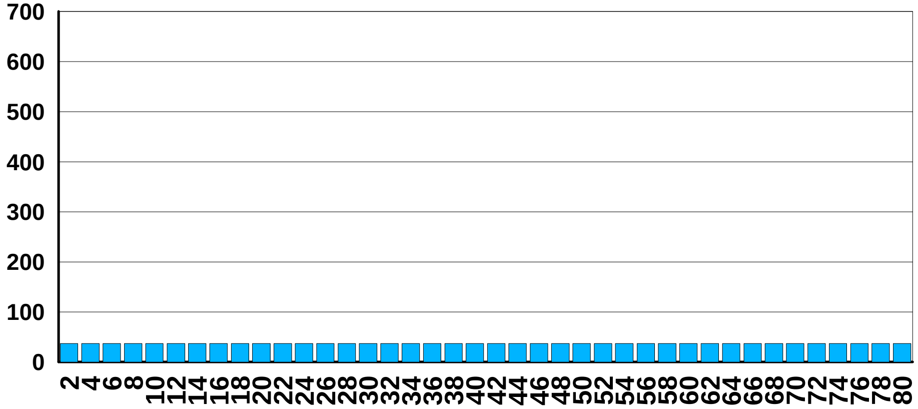
Load Distribution Reference

x axis = # of guests
y axis = transactions per second

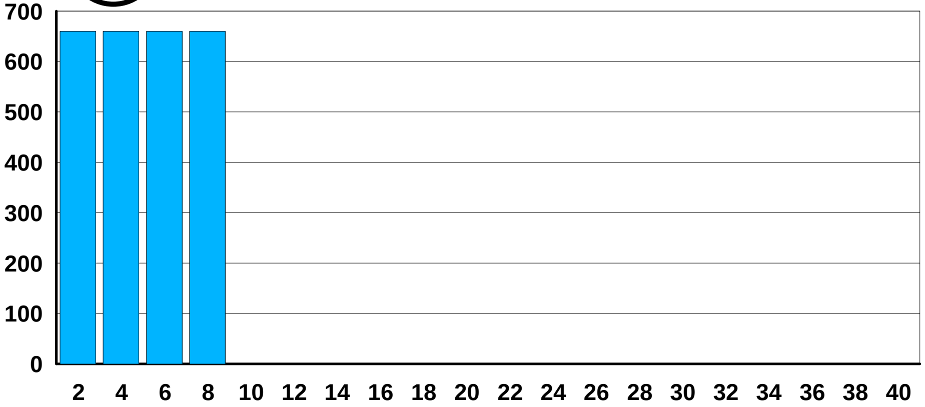
A Stress 20 streams



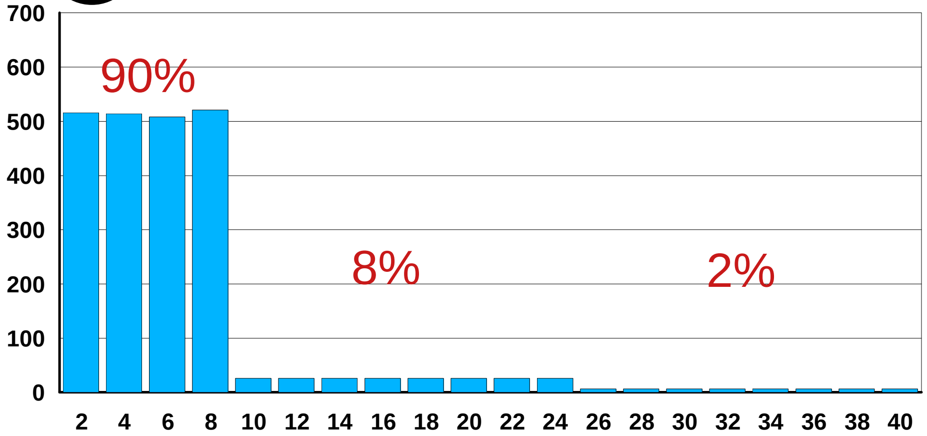
B Stress 40 streams



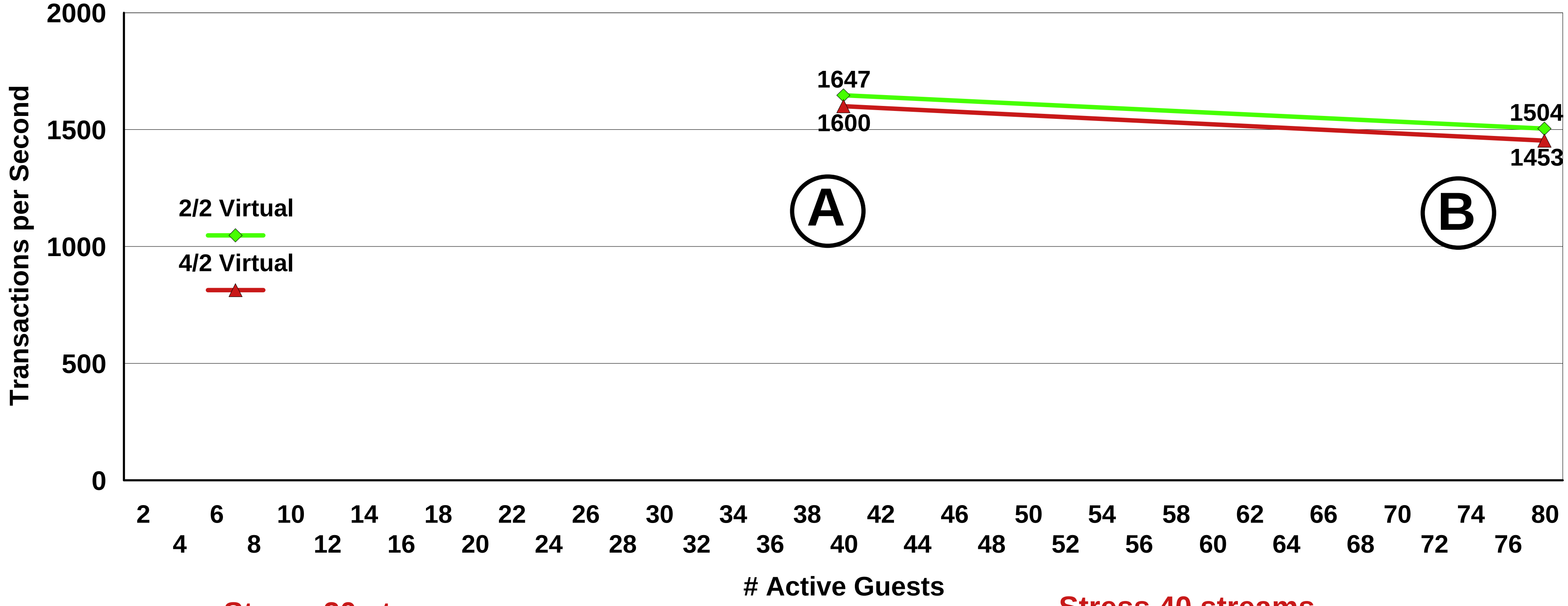
C Operational - 4 busy, 20 streams



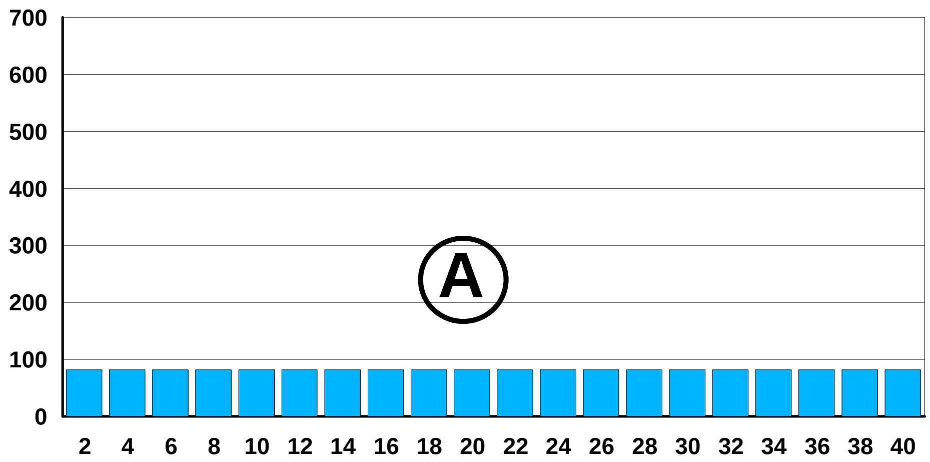
D Operational - 90/10 skew, 20 streams



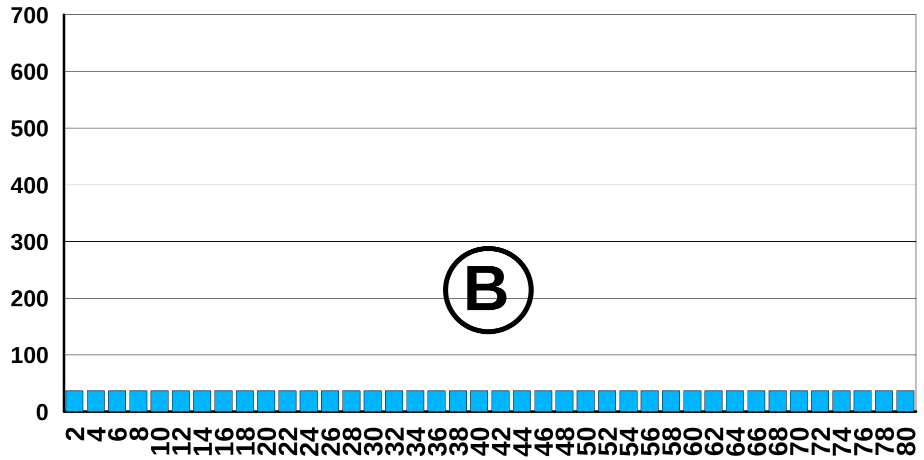
Benchmark Stress Test



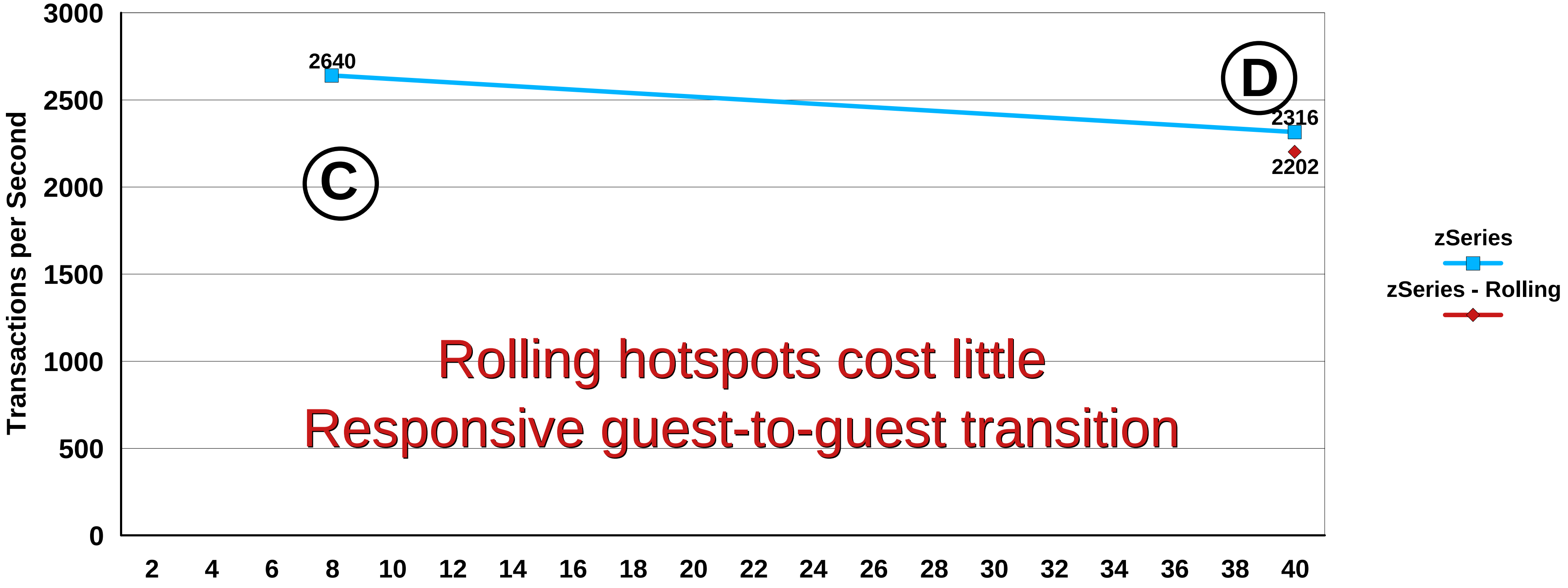
Stress 20 streams



Stress 40 streams



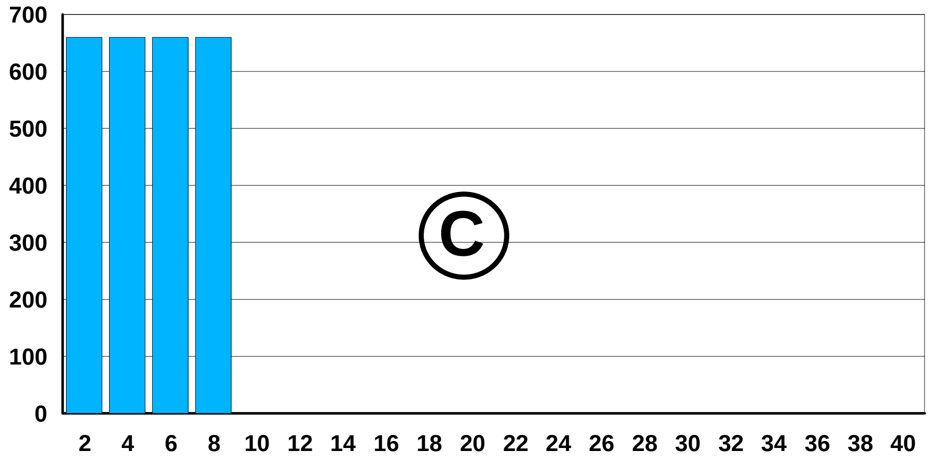
Operational Performance Test



Rolling hotspots cost little

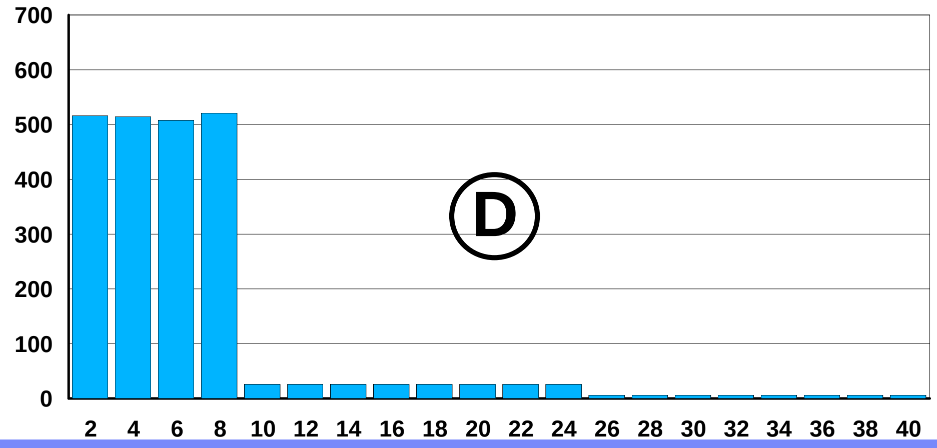
Responsive guest-to-guest transition

Operational - 4 busy, 20 streams

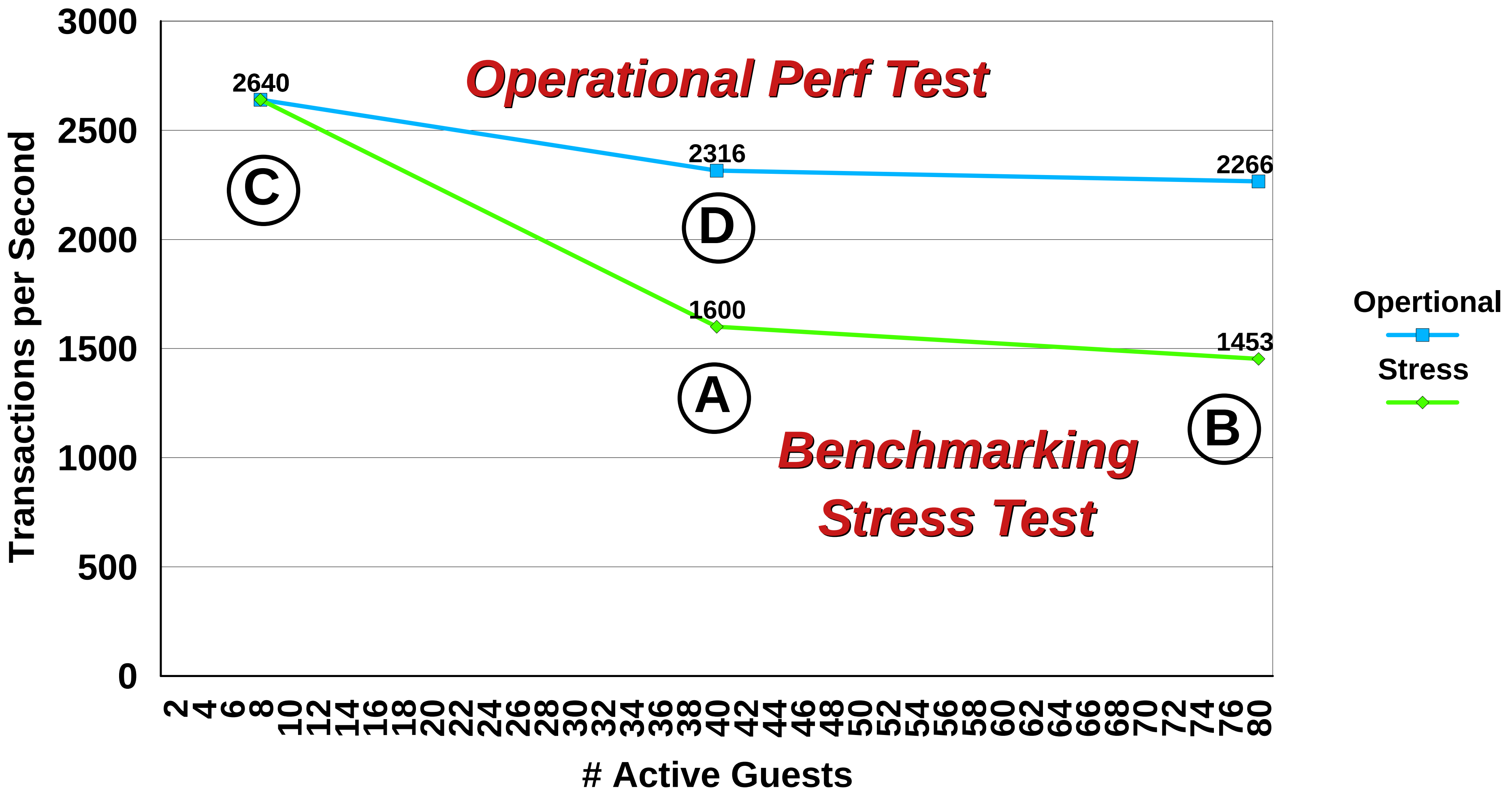


Operational - 90/10 skew, 20 streams

Static & 1 minute roll



Total System Throughput



45% increase in performance w/ real world skew

- with 40 guests - 160 virtual cpus on 16 physical cpus - 10:1 ratio

Summary

- Many features to be exploited
- The answer is "It depends. With Linux, it depends even more"
- Optimum configuration will depend on
  - What you mean by the term performance
  - What resources you have available
- See VM home page for additional information: www.vm.**ibm.com**/perf/