Large Scale Linux

Klaus Bergmann
L80
zSeries Expo, November 10 -14, 2003 | Hilton, Las Vegas, NV

Trademarks

The following are trademarks of the International Business Machines Corporation in the United States and/or other countries.
Enterprise Storage Server
ESCON*
FICON
FICON Express
HiperSockets
IBM*
IBM logo*
IBM eServer
Netfinity*
S/390*
VM/ESA*
WebSphere*
z/VM
zSeries
* Registered trademarks of IBM Corporation
The following are trademarks or registered trademarks of other companies.
Intel is a trademark of the Intel Corporation in the United States and other countries.
Java and all Java-related trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc., in the United States and other countries.
Lotus, Notes, and Domino are trademarks or registered trademarks of Lotus Development Corporation.
Linux is a registered trademark of Linus Torvalds.
Microsoft, Windows and Windows NT are registered trademarks of Microsoft Corporation.
Penguin (Tux) compliments of Larry Ewing.
SET and Secure Electronic Transaction are trademarks owned by SET Secure Electronic Transaction LLC.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* All other products may be trademarks or registered trademarks of their respective companies.

I

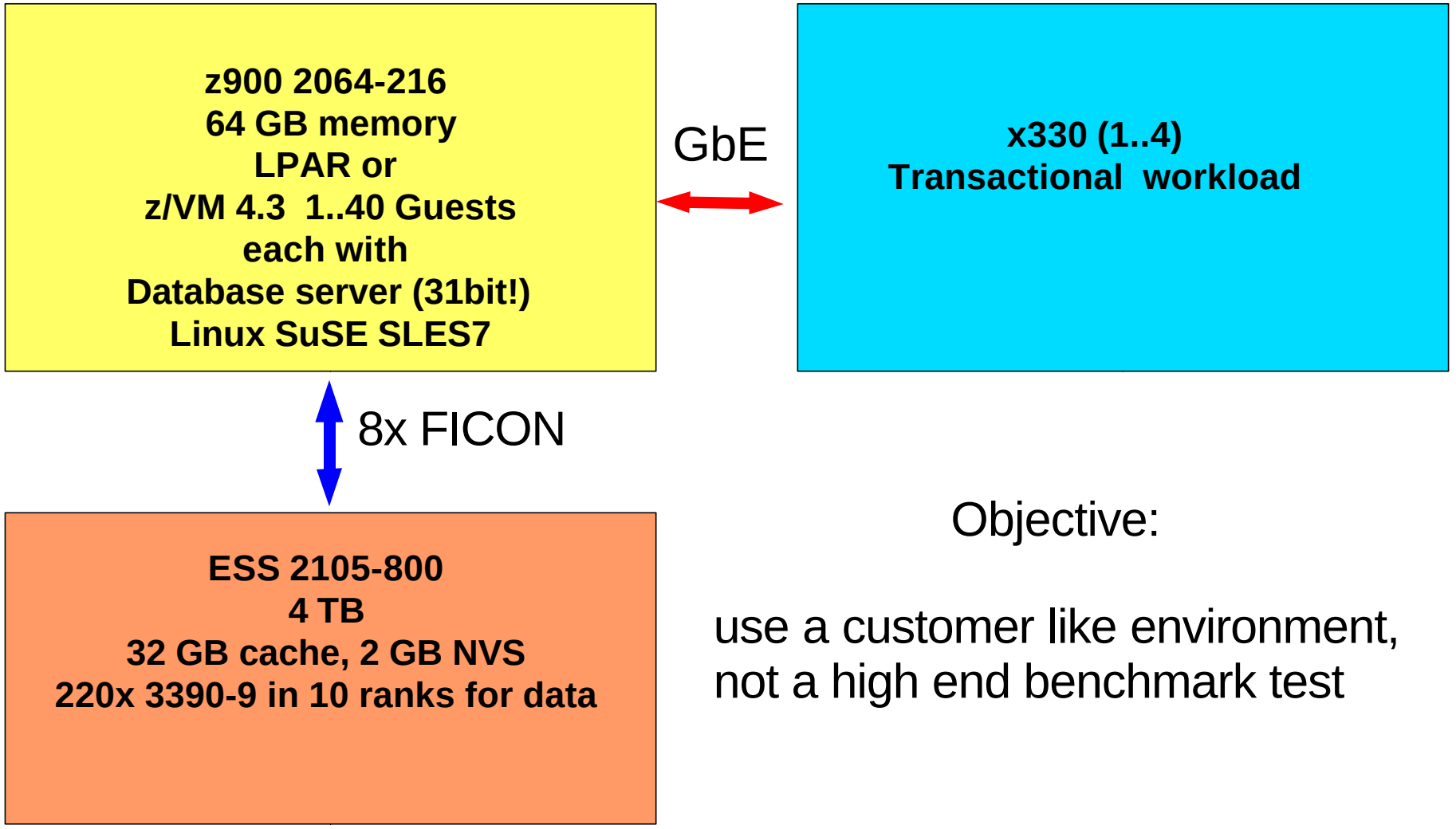
Agenda

- Experiences with database tests
  - Overview
  - Setups
  - Single Server
  - Multi Servers
- Network devices – Which one is the best for your penguin colony ?

Linux Large Scale Solution Test Center (LSC)

- Large scale horizontal and vertical solution testing of key IBM and ISV products
  - Drive configuration to the limits and above
  - Feedback to
    - Marketing/Sales
    - Sizing
    - Tech Support
    - Design & Development
  - Development of best practice implementation and tuning techniques
- Customer orientation
  - Use GA Hardware & Software (VM, Linux, Middleware, ISV, etc)
  - LPAR or VM with many guests
  - Customer like environments

Test Environment



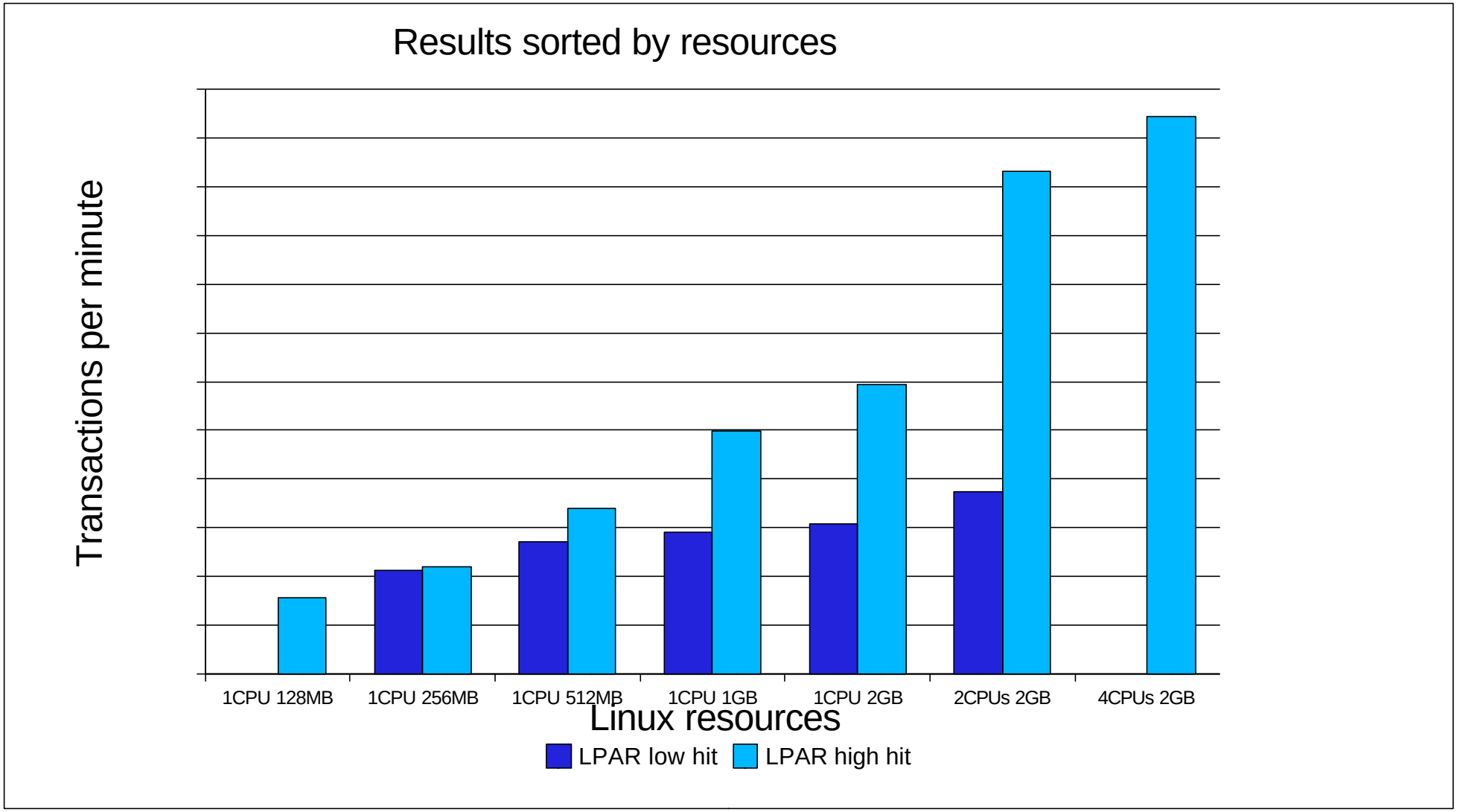
Objective:

use a customer like environment,
not a high end benchmark test

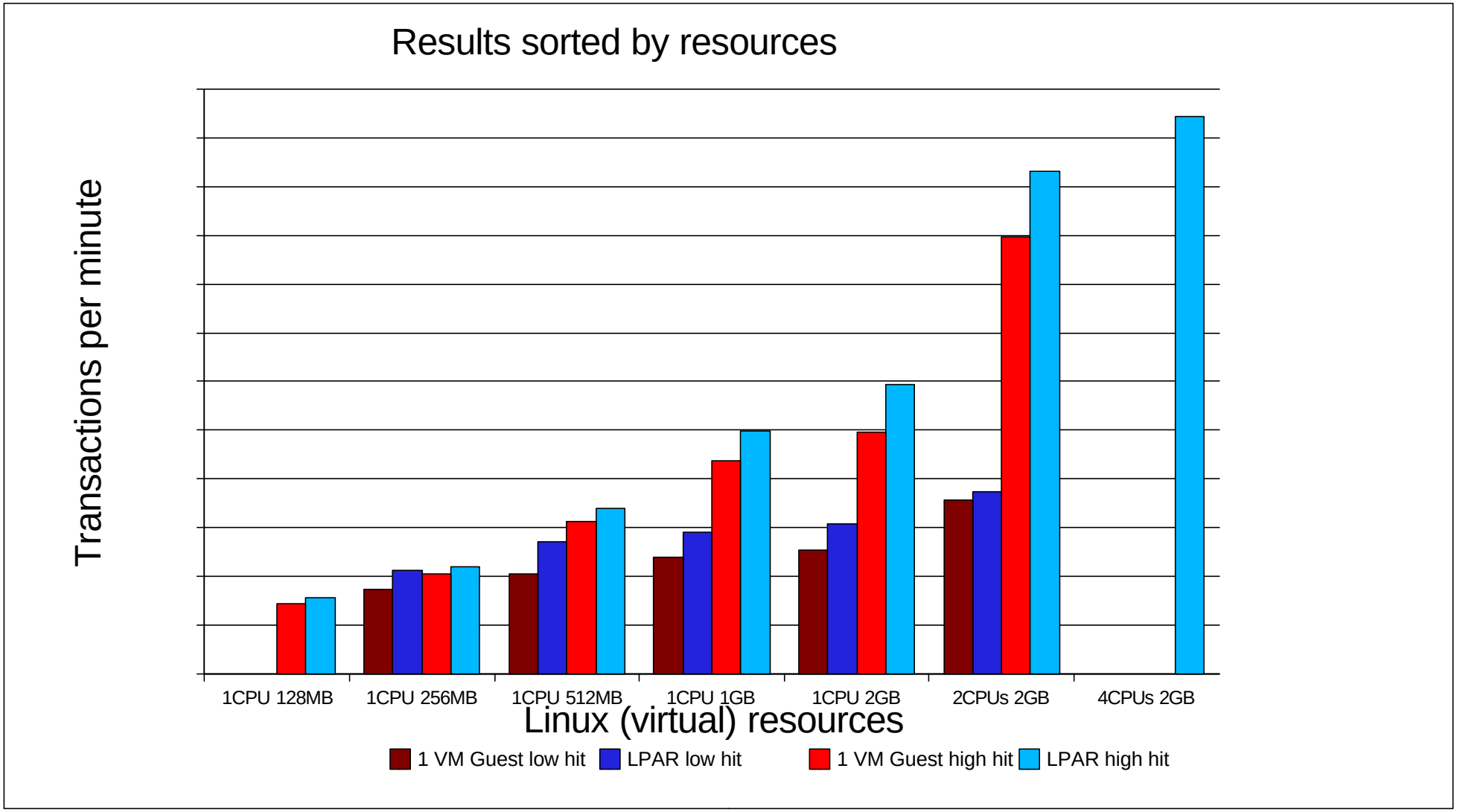
Workload description

- Transactional workload, mix of reads and writes
  - Simulates user transactions of an order-entry environment
  - Includes inquiries and updates
  - No think time / key time
  - No transaction concentrator
  - Databases up to 120 GB
  - Random access on database rows
  - Tests with 80% and >90% database buffer hit ratio

Single server results



Single server results



Single server observations

- Throughput with high hit ratio:
  - Scaling from 1 to 2 CPUs = 2x
  - Maximum difference to low hit ratio = 2.5 x
  - Memory scaling affects transaction throughput
- Throughput with low hit ratio:
  - No big difference between 1 CPU, 512 MB and 2 CPUs, 2 GB
  - Many disk accesses are needed.
  - Disk access is random, I/O requests carry 4 KB or 8 KB data
- Degradation LPAR -> VM is 6% to 24%
- VM CP overhead is 6% to 12%
- 31-bit systems can address up to 2 GB memory. Maximum shared memory is 1 GB in SuSE SLES 7.

Single server performance recommendations

- Make the Linux shared memory as large as possible
  - SuSE SLES7 = 800 MB
  - SuSE SLES8 = 1.5 GB
- Linux default settings for semaphores, max. file handles, max. number of processes have to be set according to database performance recommendations
- The database disks should be spread over many ranks.
  - The transaction throughput can be improved by using disks in 10 ranks compared to a setup with all disks in 1 rank up to 4x.
- Use “normal I/O” for database disks in Linux DASD driver instead of the default “sequential I/O”.
  - The performance improvement is up to 20%. This policy can be set with SuSE SLES 8. (SuSE SLES 8 later release “tunedasd”)

Shared Memory

Kernel parameter	Default	Our settings	Purpose
SHMMAX	33554432	1073741824	Defines the max. allowable size of one shared memory segment
SHMMNI	4096	8000	Defines the max. number of shared memory segments in the entire system.
SHMALL	2097152	262144	Defines the max. shared memory system wide in pages.
SHMMIN	1	1	Defines the min. allowable size of a single shared memory segment.
SHMSEG	4096	4096	Defines the max. number of shared memory segments one process can attach.

The commands:

```
echo          8000 > /proc/sys/kernel/shmmni
echo        262144 > /proc/sys/kernel/shmall
echo 1073741824 > /proc/sys/kernel/shmmax
```

enter the appropriate values into the Kernel parameters.

/etc/sysctl.conf can also be used

Semaphores

Kernel parameter	Default	Our settings	Purpose
SEMMSL	250	100	Defines the minimum recommended value, for initial installation only.
SEMMNS	256000	32000	Defines the max. semaphores on the system. This setting is a minimum recommended value, for initial installation only. The SEMMNS param. Should be set to the sum of the PROCESSES parameter for each database, adding an additional 10 for each database.
SEMOPM	32	100	Defines the maximum number of operations for each semop call.
SEMMNI	1024	100	Defines the maximum number of semaphore sets in the entire system.

The command:
echo 100 32000 100 100 > /proc/sys/kernel/sem

enters the appropriate values into the Kernel parameters.

/etc/sysctl.conf can also be used

Our profile

/etc/init.d/boot.local:

```
# semaphore parameter values
# cat /proc/sys/kernel/sem:      250       256000          32        1024
# SEMMSL_value SEMMNS_value  SEMOPM_value SEMMNI_value
echo 100 32000 100 100 > /proc/sys/kernel/sem

# maximum shared segmant size in bytes, default is SHMMAX=33554432
# cat /proc/sys/kernel/shmmax:       33554432
echo 1073741824 > /proc/sys/kernel/shmmax

# maximum number of shared segments system wide, default is SHMMNI=4096
# cat /proc/sys/kernel/shmmni:       4096
echo 8000 > /proc/sys/kernel/shmmni

# maximum shared memory system wide in pages, default is SHMALL=2097152
# cat /proc/sys/kernel/shmall: 262144
echo 262144 > /proc/sys/kernel/shmall
```

Our profile, cont.

cat /proc/sys/fs/file-max: 8192
echo 65536 > /proc/sys/fs/file-max
ulimit -n 65536

Set the Sockets to /proc/sys/net/ipv4/ip_local_port_range
cat /proc/sys/net/ipv4/ip_local_port_range: 32768   61000
echo 1024 65000 > /proc/sys/net/ipv4/ip_local_port_range

Set the Process limit by using *ulimit -u*.
This will give you the number of processes per user.
ulimit -u 16384

Set new eth parameter USE MTU Size: 8992 for jumbo frames
/etc/rc.config:
IFCONFIG_1="10.0.0.200 broadcast 10.0.255.255 netmask 255.255.0.0 mtu 8992 up”

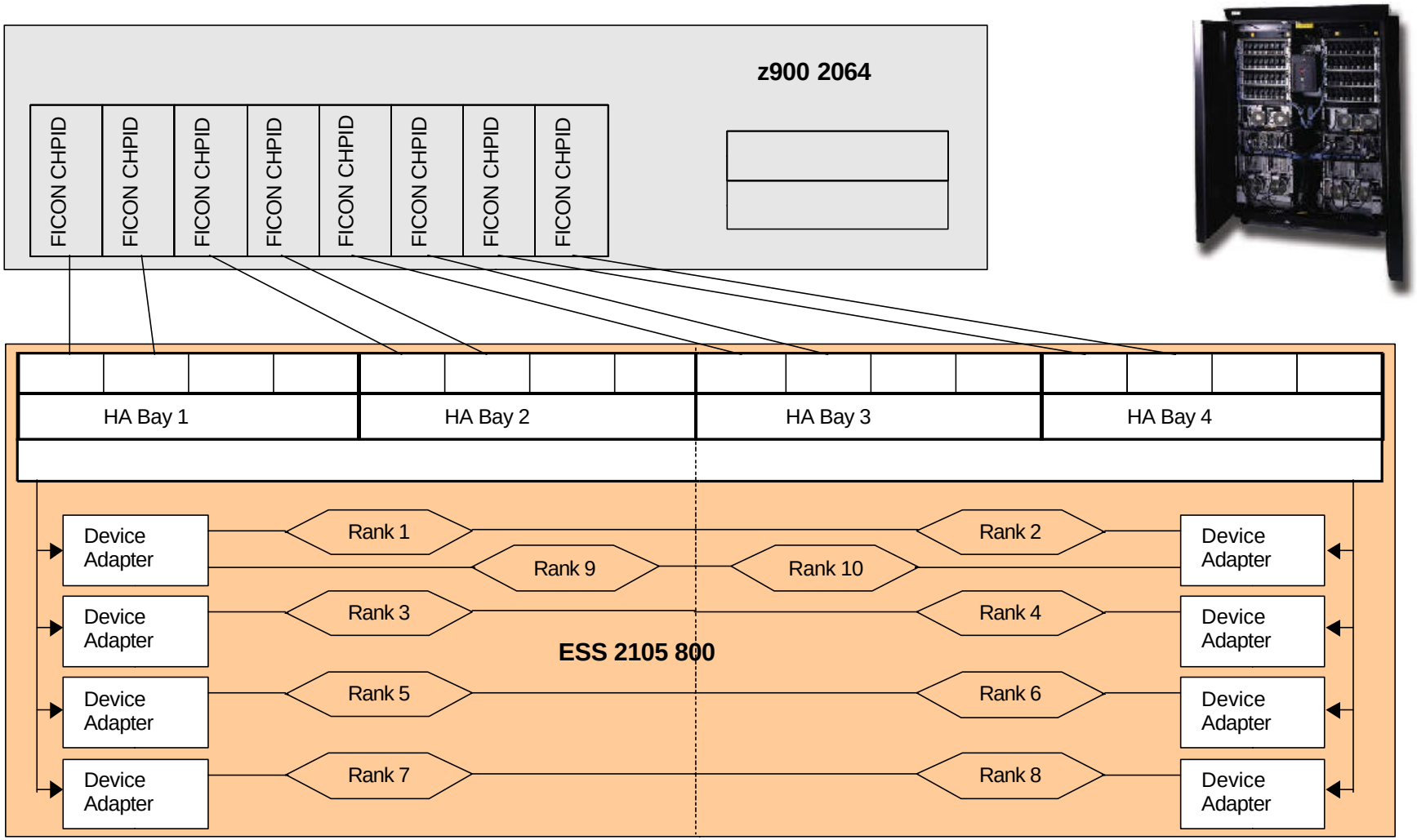
VM setup for many server test

CPUs	8
MEMORY	15 GB central
XSTORE	4GB, the default recommendation of 2GB could not handle the large amounts of database disk I/Os.
PAGE DEVICES	4x 3390-3 in different ranks, the test was run so that only little paging activity occured
SET MDC SYSTEM OFF	Minidisk cache is a read cache. The random nature of the workload did not benefit from minidisk cache
Minimum TIMESLICE	The default of 5ms worked acceptable for up to 8 guests. 20 or 40 guests needed longer timeslices (25 ms)

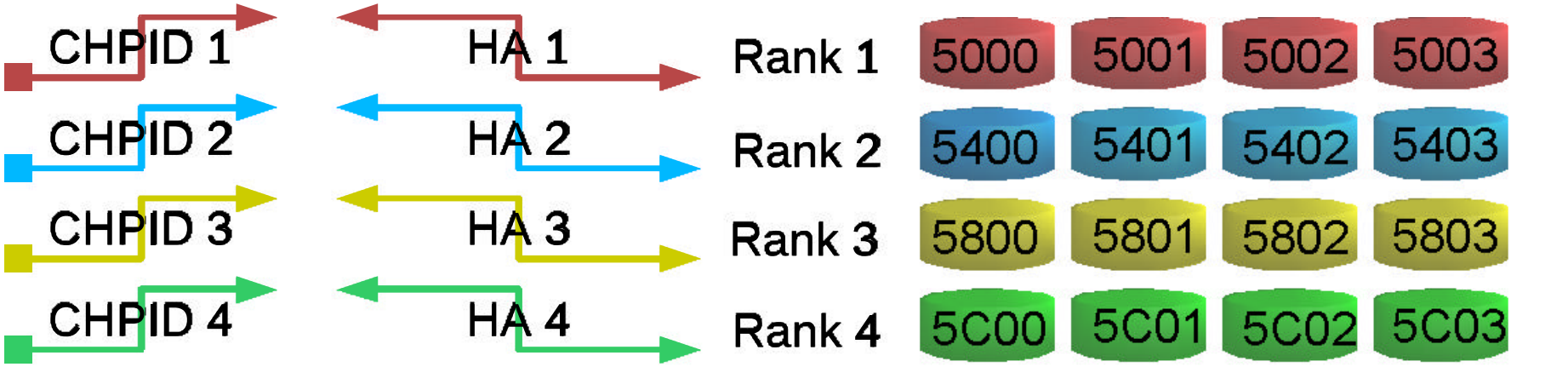
VM guest setup

CPUs	Use 1 virtual CPU unless your Linux guest urgently needs more CPUs to get the usual work done.
MEMORY	Use minimum amount of memory for your Linux guest. Find limit, where swap begins. Remember that Linux uses always all of its memory. VM then estimates working set too large. Different setups used 1 GB, 384 MB, 256 MB and 144 MB
MINIDISK or DEDICATED?	I/O throughput is identical for fullpack minidisks and dedicated disks. In the test we used minidisks for the Linux installations because they can be shared among guests (cloning), and dedicated disks for the database tables. 8 guests setup: 22x 3390-9 per server 40 guests setup: 4x 3390-9 per server
ABSOLUTE SHARE	Tests with many active database servers showed that the setting of absolute share for a few servers did not improve their performance, because this option can only help if CPU is the bottleneck
QUICKDSP	= ON is justified only for a small number of guests Many guests should use OFF

Disk configuration



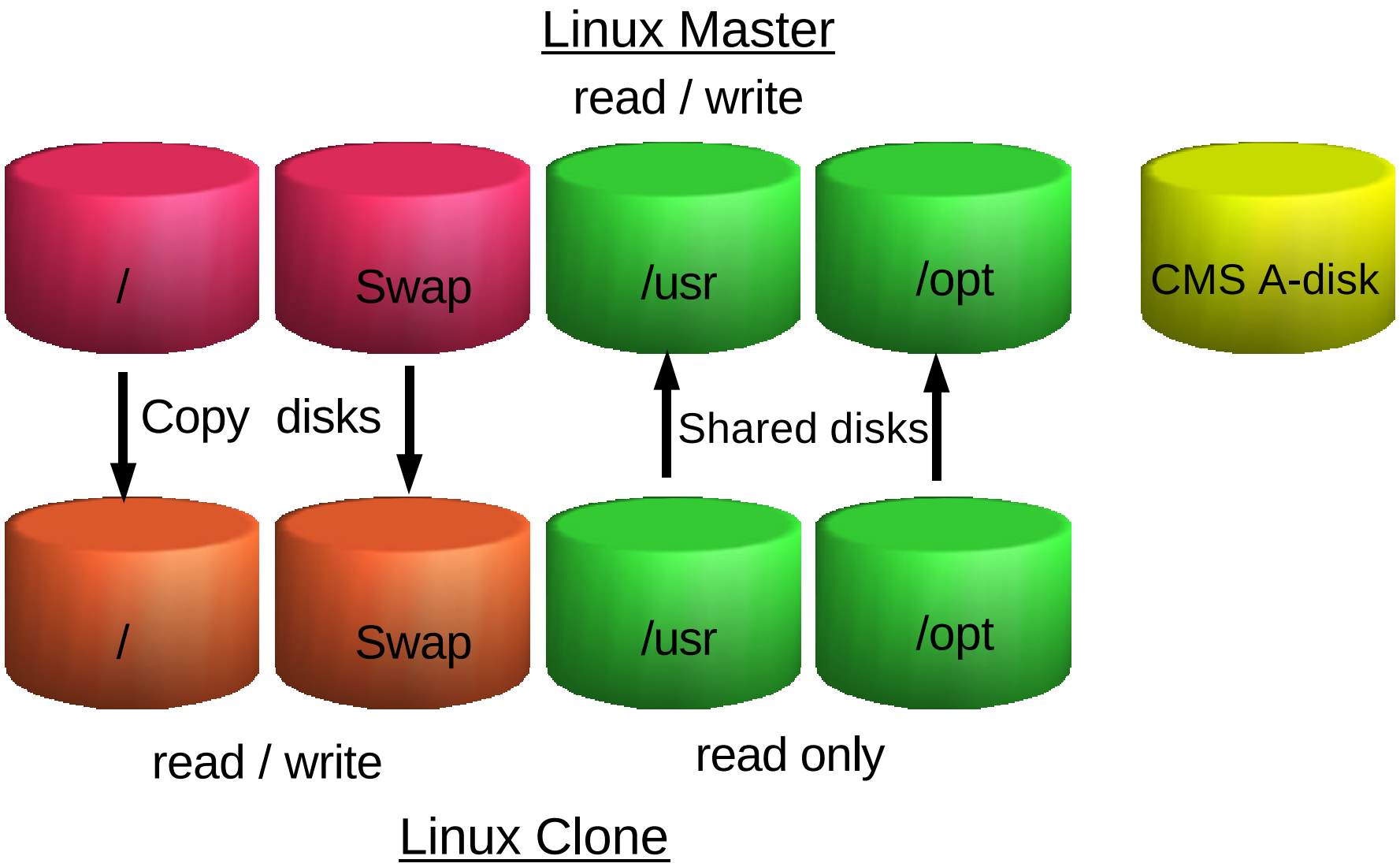
VM guest disk usage



Define LVs:
with stripes 32KB/64KB

Define z/VM guests:

VM guest cloning



VM guest customization

Linux Master

IPL Linux
mount / of Linux Clone

Customize each guest

hostname

ip address

/etc/fstab

/etc/chandev.conf

/boot/parmfile

SuSE SLES7 rc.config

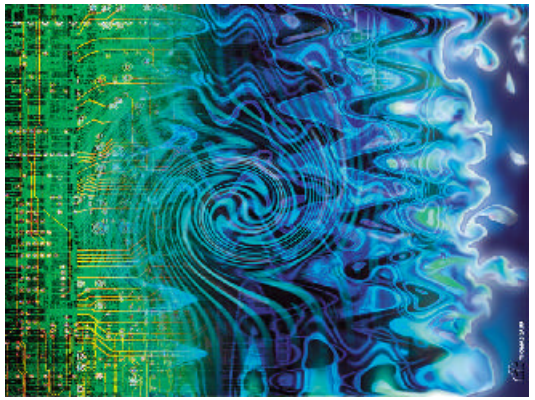
zipl

Linux Clone

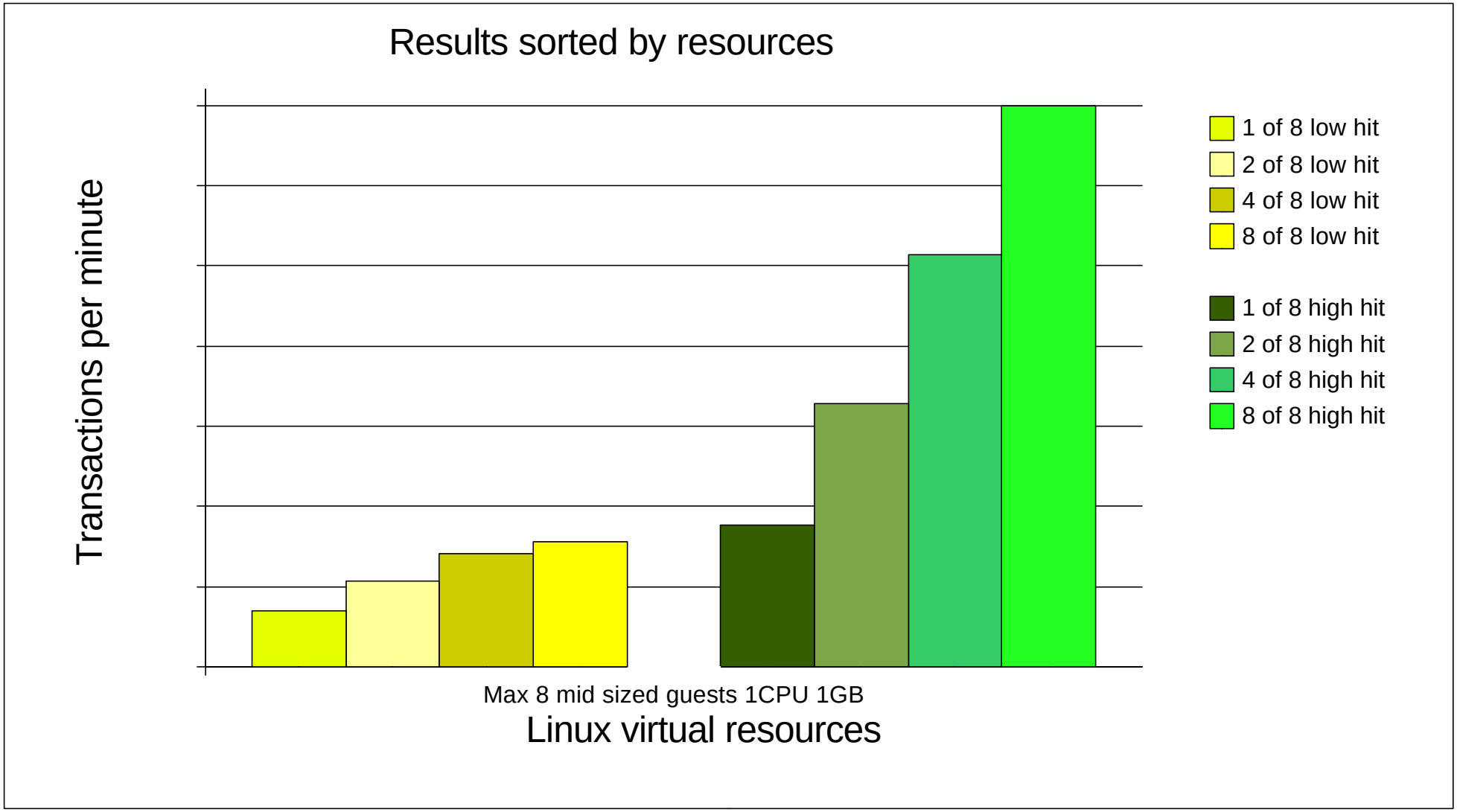
Read / write

Multi servers test setup

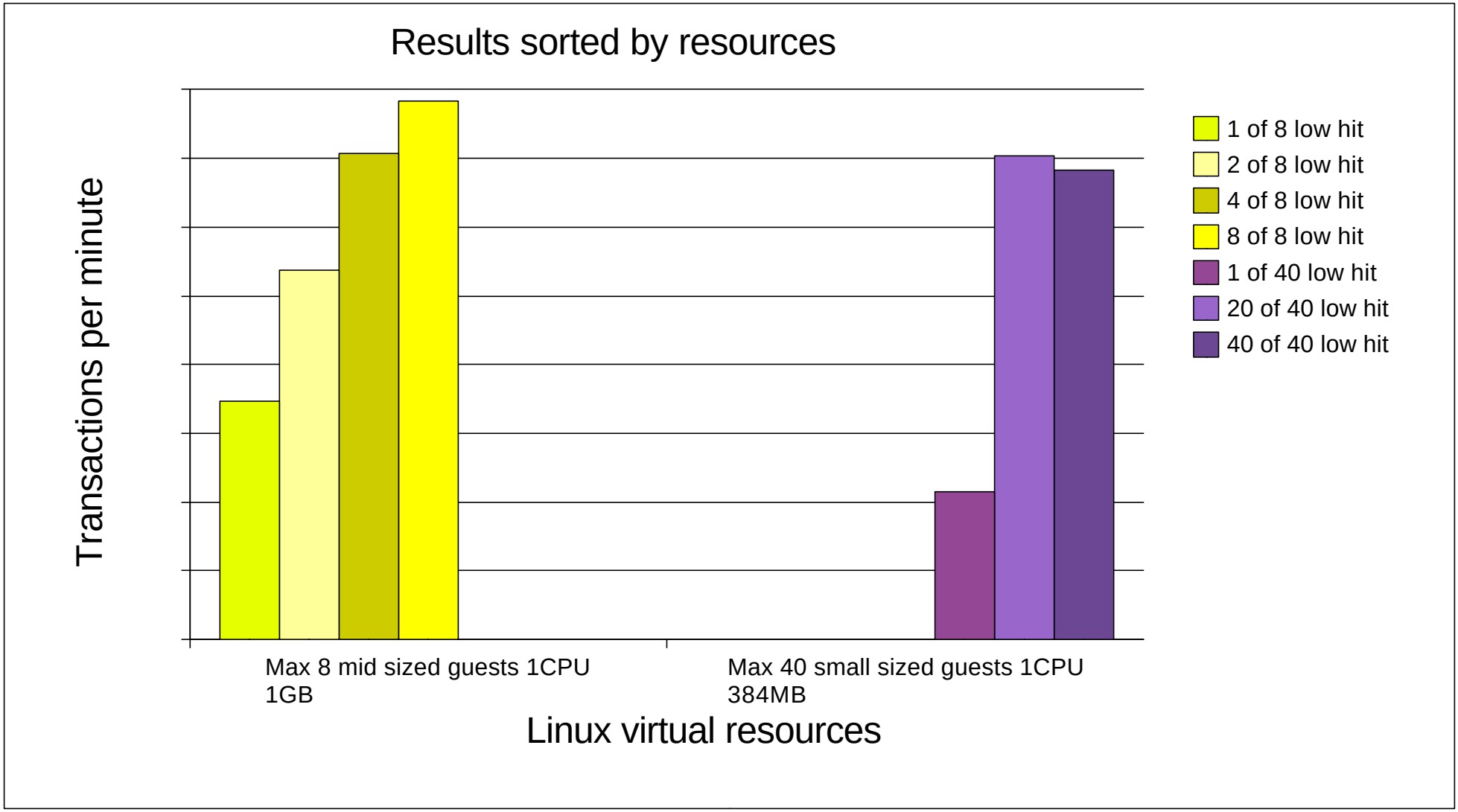
- 8 mid sized database servers:
  - 1 virtual CPU, 1 GB memory, 22x 3390-9 disks for database tables
- 40 small sized servers, balanced workload:
  - 1 virtual CPU, 384 MB memory, 4x 3390-9 disks for database tables
- No idle servers !
  - This does not reflect real production environments



Results with 8 mid sized servers



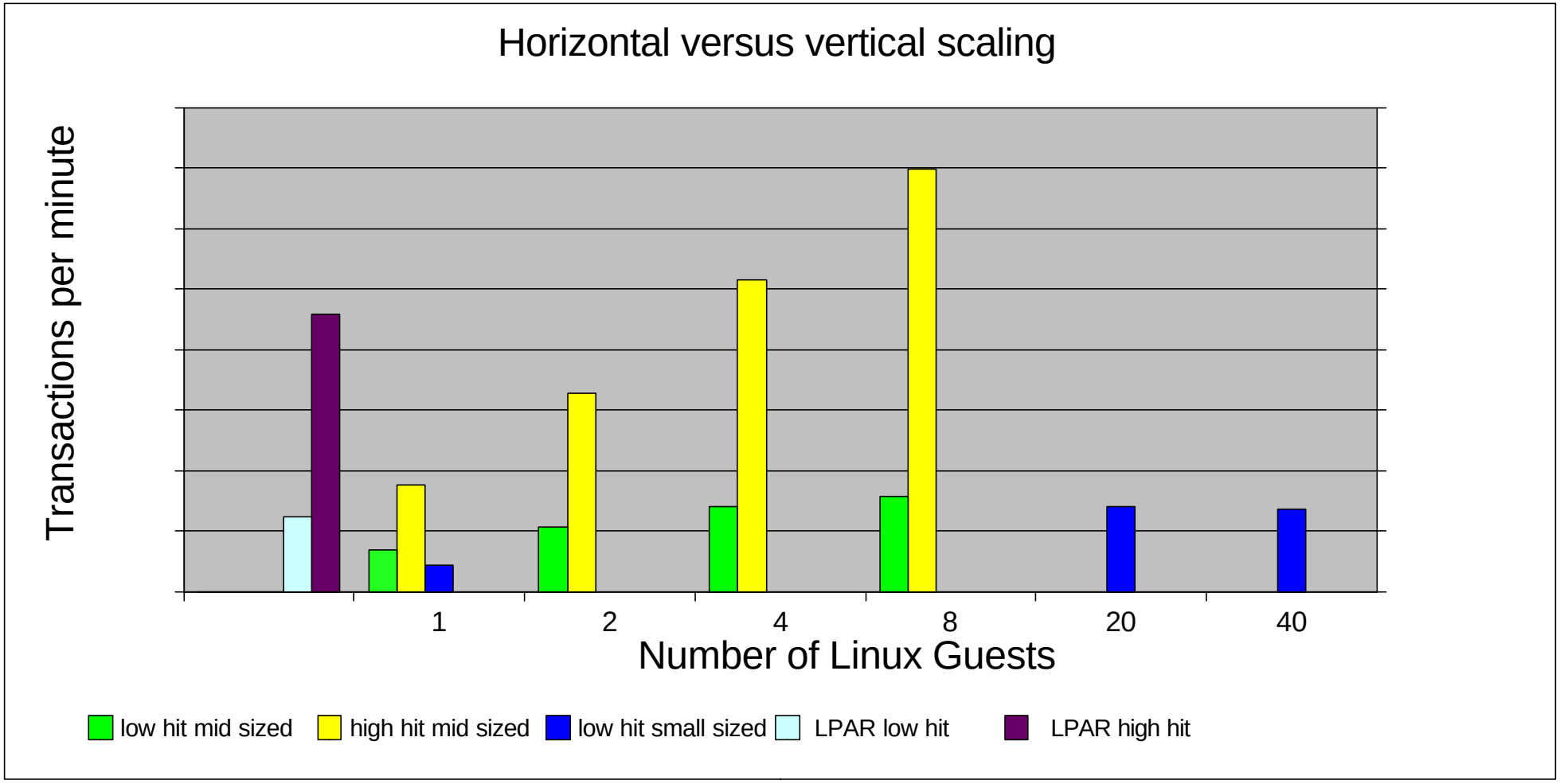
Results with up to 40 small servers



Multi servers observations

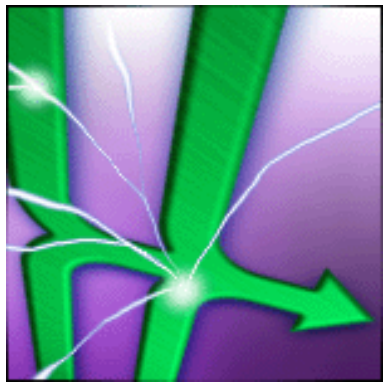
- Total number of disk I/O requests is 8000 SSCH/sec.
  - A storage server in a production environment usually runs at 3000 – 5000 SSCH/sec.
- With low hit ratio the performance of 40 small sized servers and 8 mid sized servers is almost identical.

Multi-server versus single-server results



Many servers versus single server observations

- High hit ratio
  - 8 mid sized servers perform better than one big single server (1.5x)
- Low hit ratio
  - 40 small sized database servers perform almost identical as 8 mid sized servers or a single big server.



Multi servers performance recommendations

- Recommendations for the single server.
- Provide a big XSTORE in VM (4 GB+).
- 64-bit databases will allow bigger single servers to reach good database buffer hit ratios and reduce high I/O loads.
- Size the Linux guests' memory carefully:
  - Don't give room to buffer cache.
  - There should be little swapping activity in the Linux guest.
  - VM can handle I/O requests from guests better if the “I/O areas” of the guests are small.
- If transaction response time is bad (low database buffer hit ratio?), increase memory and shared memory size of the database server.
- In scenarios with many busy servers:
  - Don't specify QUICKDSP ON
  - Increase the TIMESLICE from 5ms to a higher value (25ms)
  - Modifying share options of a single guest does not help when the overall disk I/O rate is high

Conclusion

- Single servers can use up to 4 CPUs.
- Few database servers under VM can drive a higher total load than a single server.
- Newer Linux distributions can provide larger shared memory than SuSE SLES7.
- 64-bit databases will allow bigger single servers to reach good database buffer hit ratios and reduce high I/O loads.
- Redbook Recommendation:
  “Experiences with Oracle for Linux on zSeries”
  SG24-6552, 4/2003
  “e-Business Intelligence: Leveraging DB2 for Linux on S390”
  SG24-5687, 7/2001

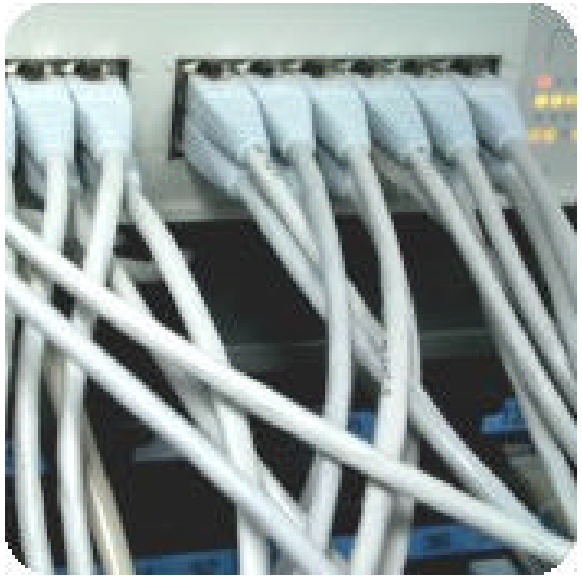
64-bit database status

Source: zLinux Mailing List at Marist,
http://www2.marist.edu/htbin/wlvindex?linux-vm

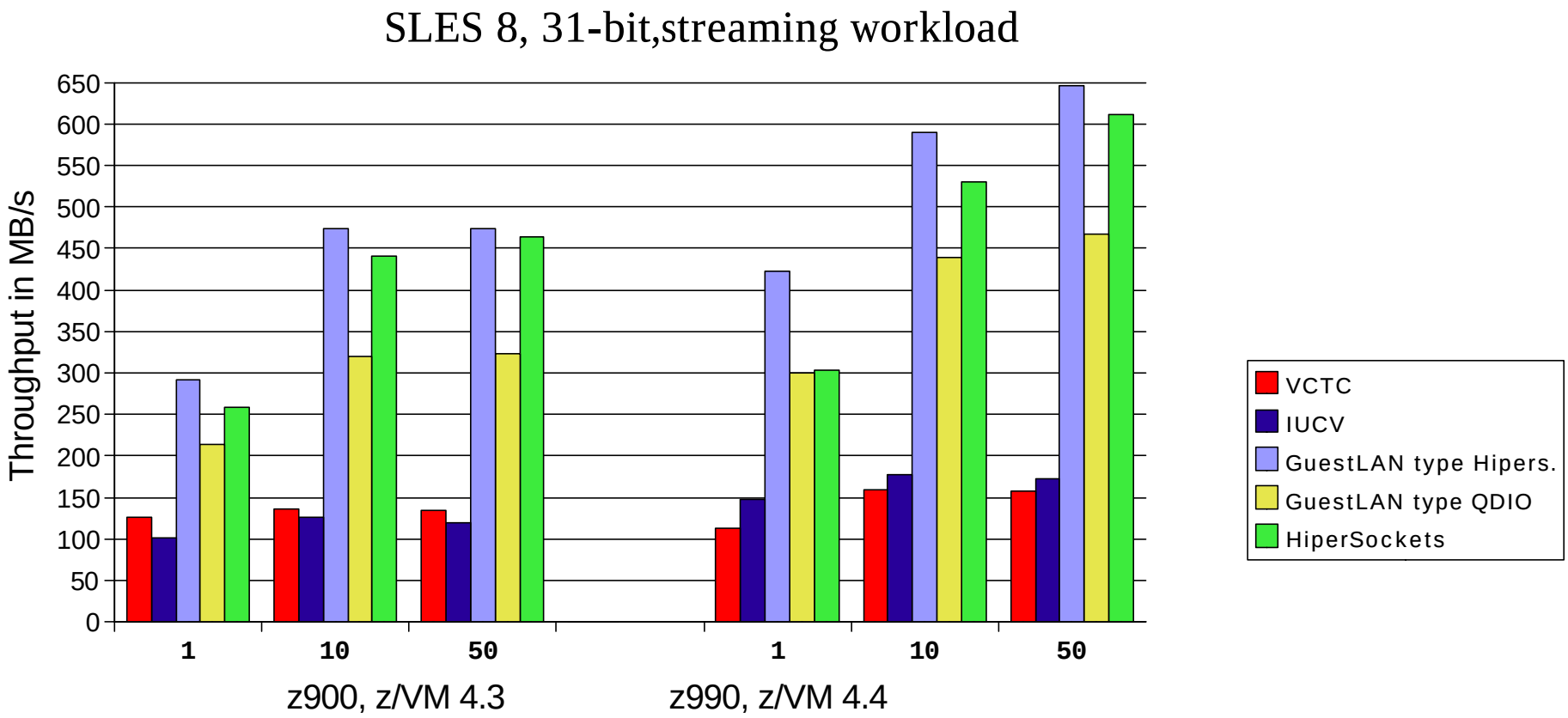
- DB2:
  “IBM plans to offer a 64-bit DB2 for Linux on zSeries. I can't say when this is coming, but it is a priority for us.”
  Jim Elliott, 10/22/03
- Oracle:
  “Oracle is in the process of certifying 31-bit 9i to run on 64 bit linux. 10g will be 64 bit only.”
  Chris Little, 10/20/03
- “FYI, Informix IDS 9.40 64-bit version is already available for z/Linux (SuSE SLES8).”
  Andreas Breitfeld, 10/23/03

Network devices

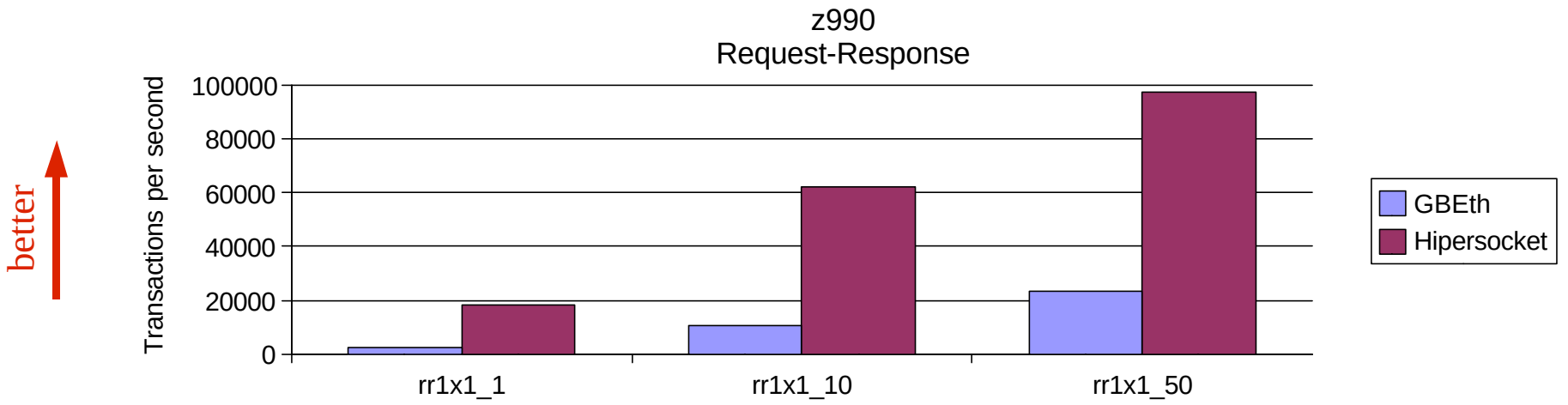
Which one is the best for your penguin colony ?

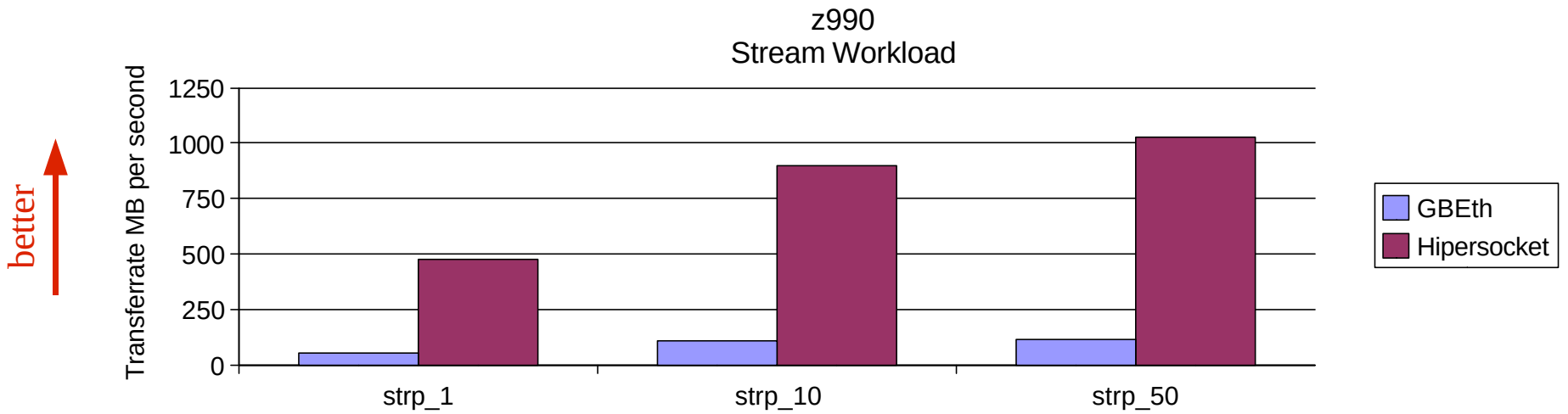


Networking for your penguin colony



Gigabit Ethernet MTU 9000/Hipersocket MTU 32K – LPAR





Which network device should I use ?

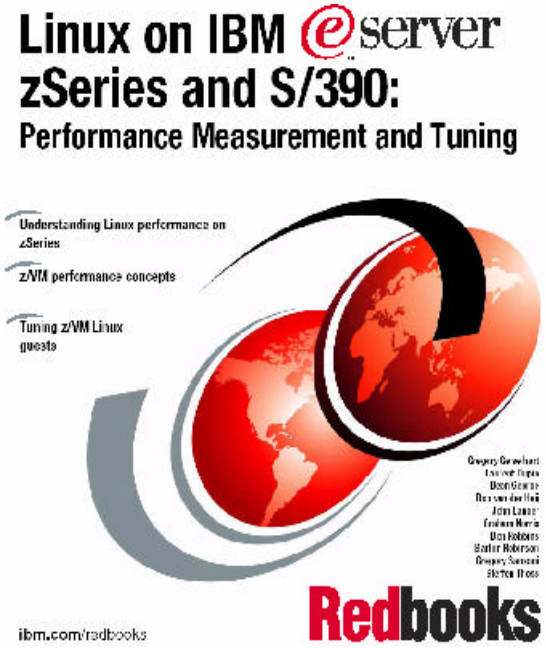
- Use GuestLAN type HiperSocket for inter-z/VM guest connections
  - performance comparable to HiperSockets
  - easy to use
  - usable on machines older than z800/z900 (z/VM 4.3. req.)
  - More connections possible than with HiperSockets
- If Multi- and Broadcasts are necessary in your z/VM environment use GuestLAN type QDIO
  - performance slightly below GuestLAN type HiperSocket
  - has packing capability
  - Thin Interrupt (QIOASSIST) available with z/VM 4.4

Which network device should I use ? (Cont.)

- If your system is low on memory use VCTC or IUCV
  - because each QDIO device (HiperSockets, GuestLAN) requires up to 8 MB fixed main memory
- A z/VM guest does not drop from queue Q3 if it uses a QDIO device or CTC device (APAR 63282)
  - apply PTF UM30888 on z/VM 4.3. or UM30889 on z/VM 4.4

Visit Us

- Linux for zSeries Performance Website:
  http//www.ibm.com/developerworks/opensource/linux390/perf_hints_tips.shtml
- Linux-VM Performance Website: http://www.vm.ibm.com/perf/tips/linuxper.html
- Performance Redbook:
  - SG24-6926-00



Questions