SHARE
Technology · Connections · Results

Performance Aspects of a Penguin Colony

Steffen Thoss
08/12/2003, Session # 9391

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 file system experience

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, Middle ware, ISV, etc)
  - LPAR or VM with many guests
  - Customer like environments

Test Environment

z900 2064-216
64 GB memory
LPAR or
z/VM 4.3 1..40 Guests
each with
Database server (31bit!)
Linux SuSE SLES7

GbE

x330 (1..4)
Transactional workload

8x FICON

ESS 2105-800
4 TB
32 GB cache, 2 GB NVS
220x 3390-9 in 10 ranks for data

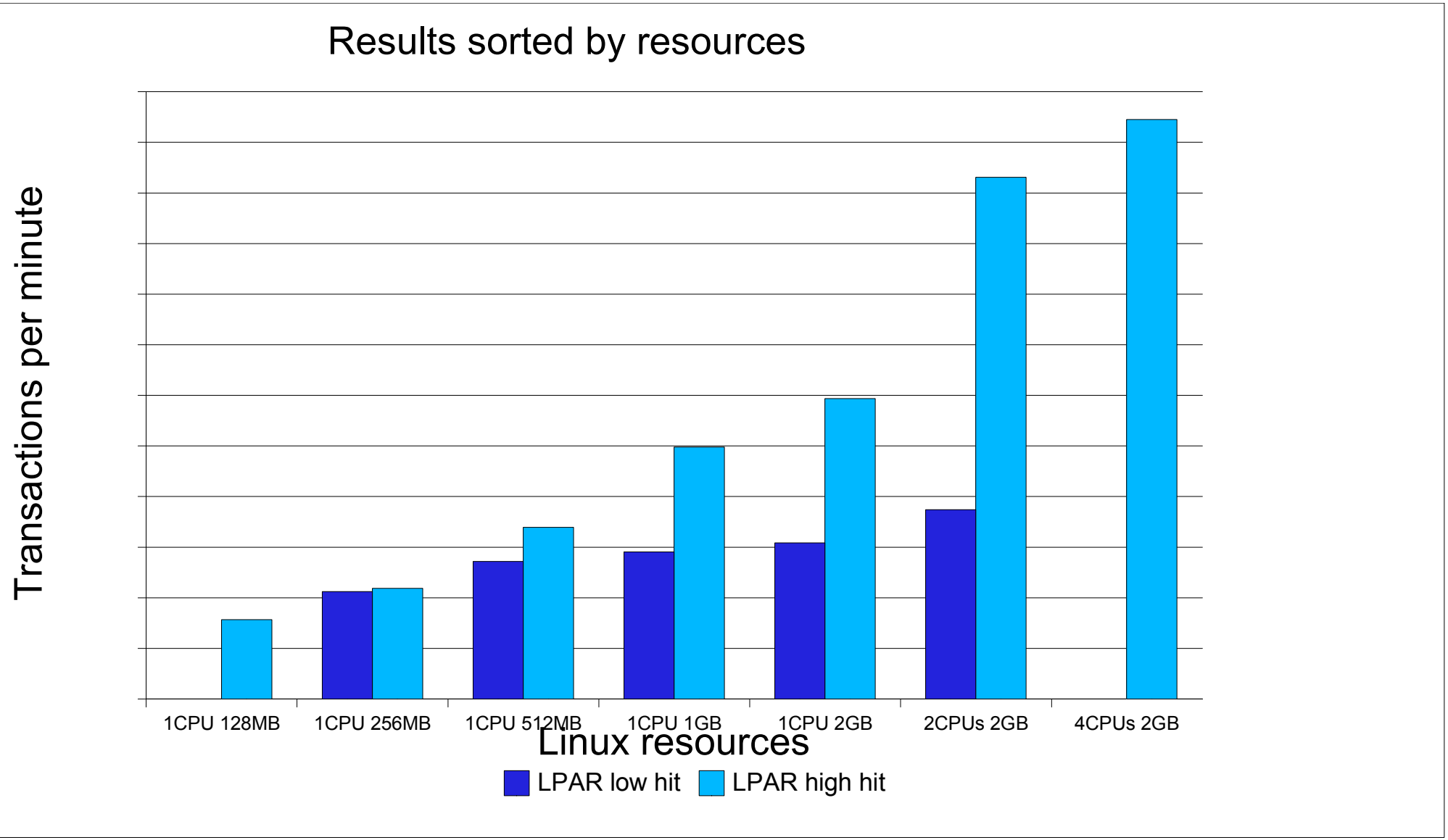
Objective:

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

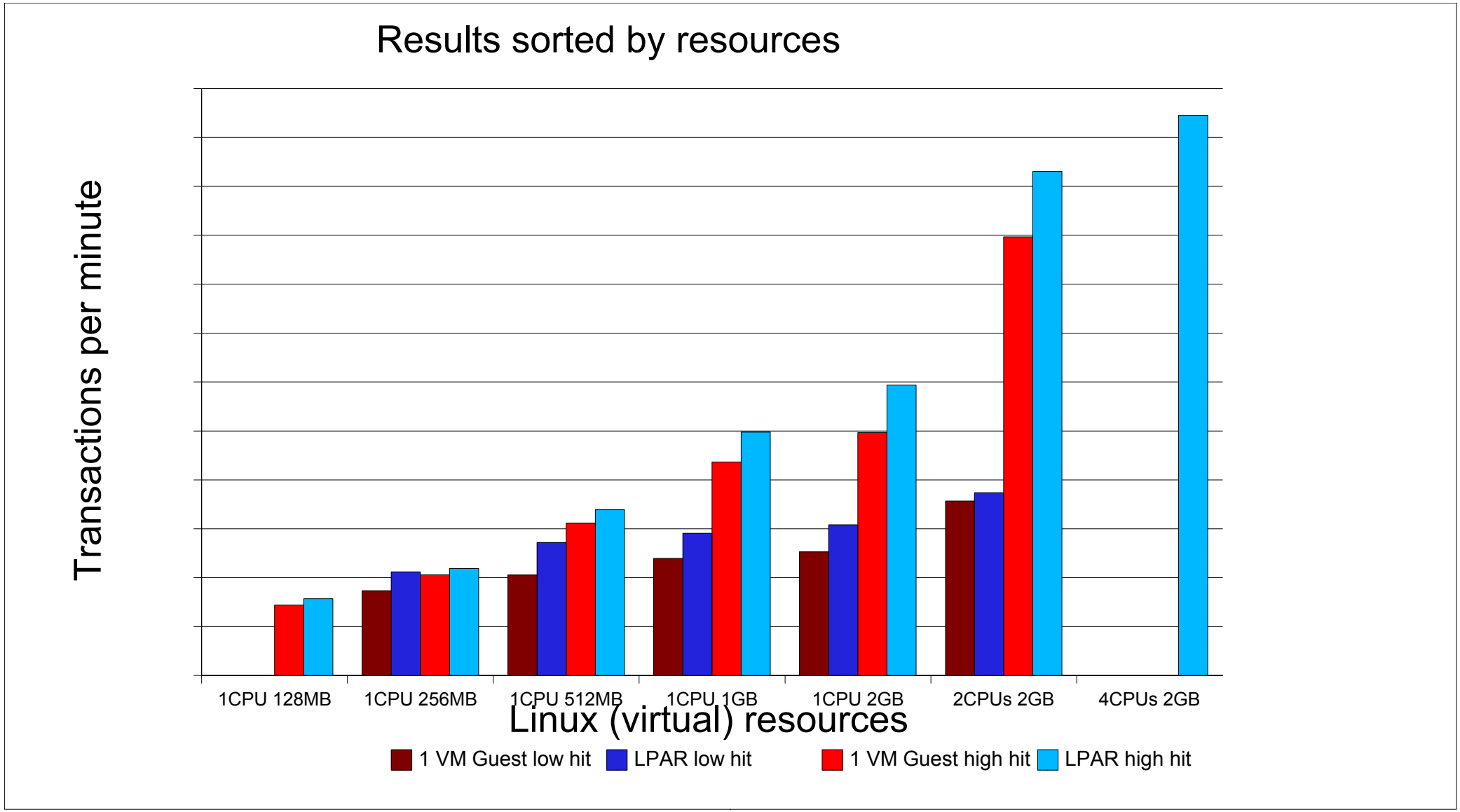
Workload description

- 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%
- 31bit 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 = 1 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”)

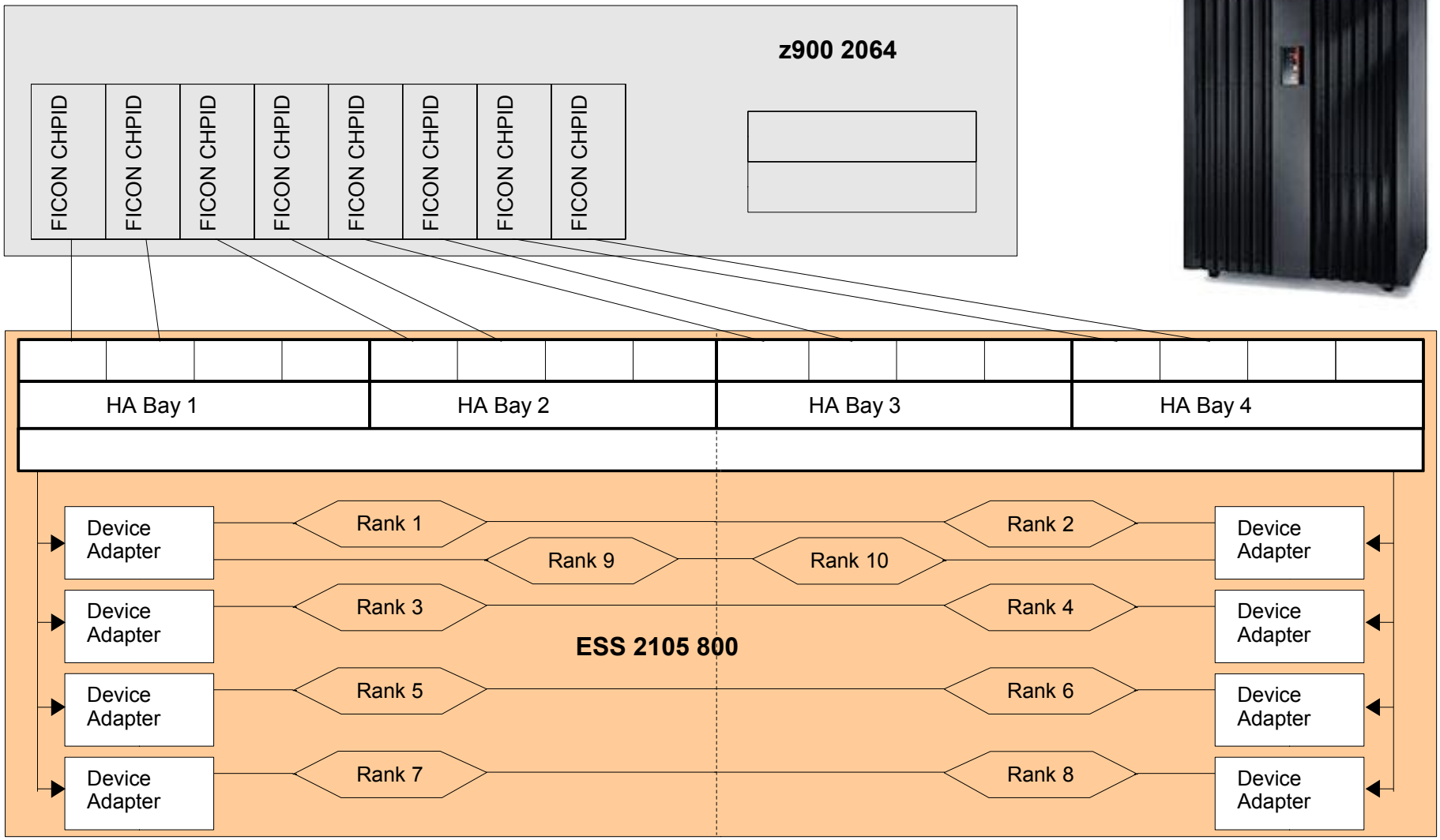
VM setup for many server test

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)

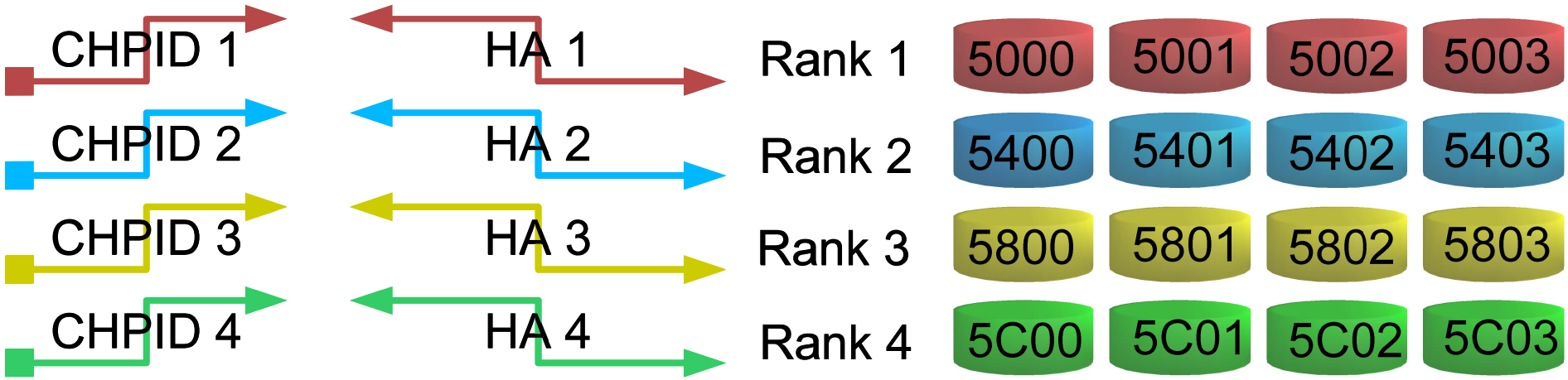
VM guest setup

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 considerable only for a small number of guests Many guests should use OFF

Disk configuration



VM guest disk usage



Define LVs:
with stripes 32KB/64KB

5000 5400 5800 5C00 5001 5401 5801 5C01 ...

Define z/VM guests:

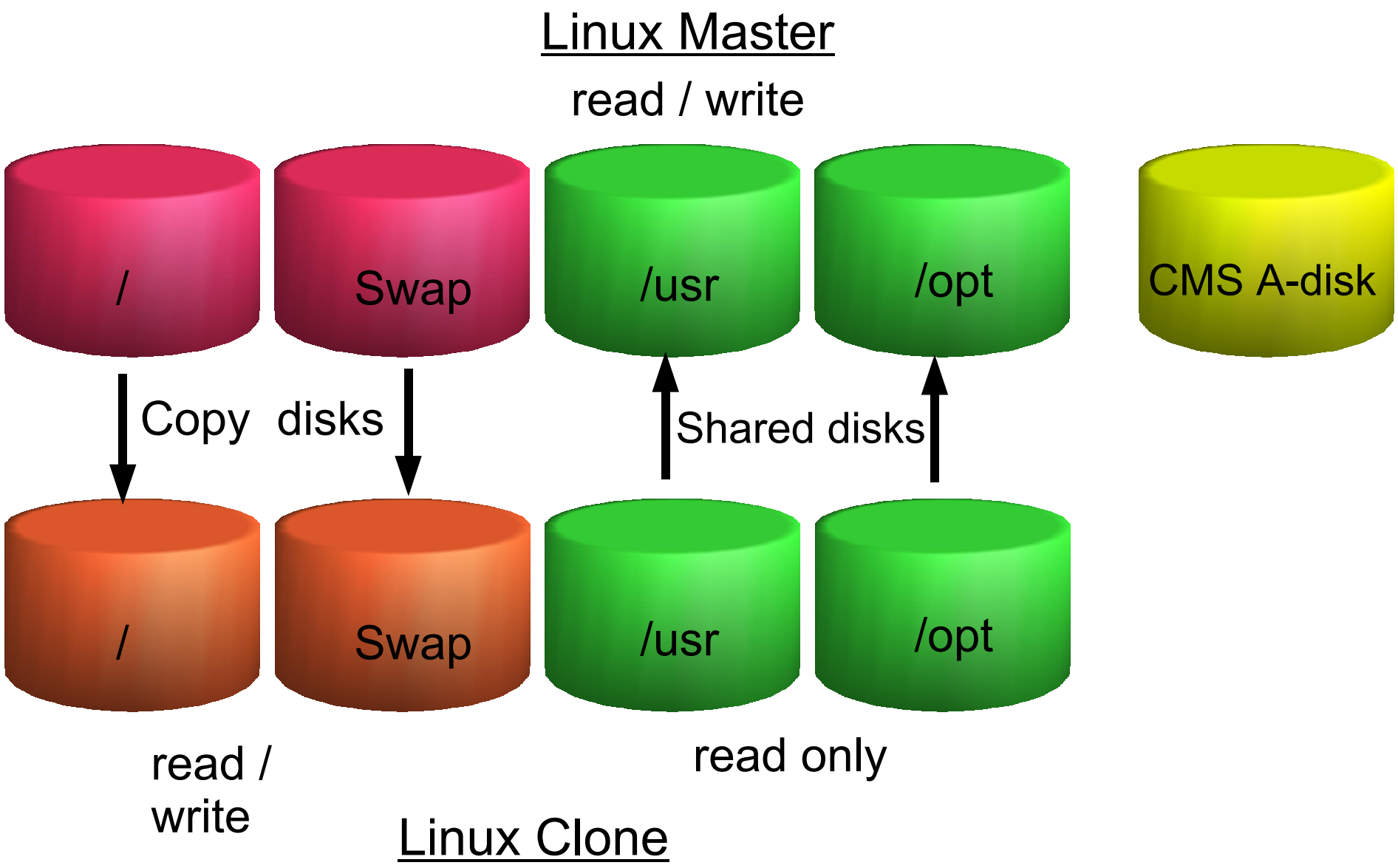
5000 5400 5800 5C00 1

5001 5401 5801 5C01 2

3 5002 5402 5802 5C02

4 5003 5403 5803 5C03

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

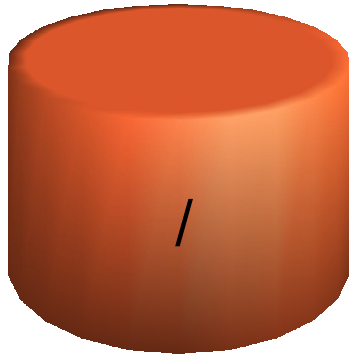
/boot/parmfile

SuSE SLES7 rc.config

zipl

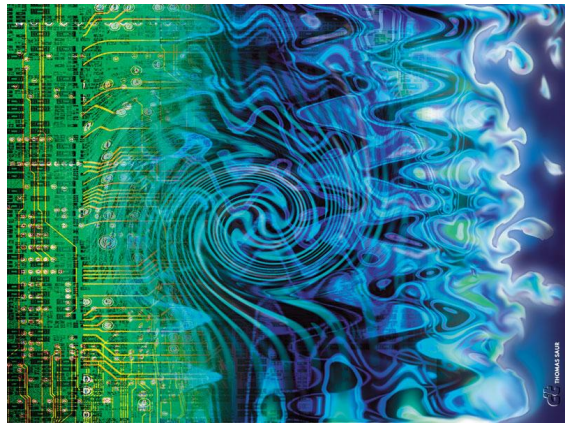
Linux Clone

Read / write



Multi servers test assumptions

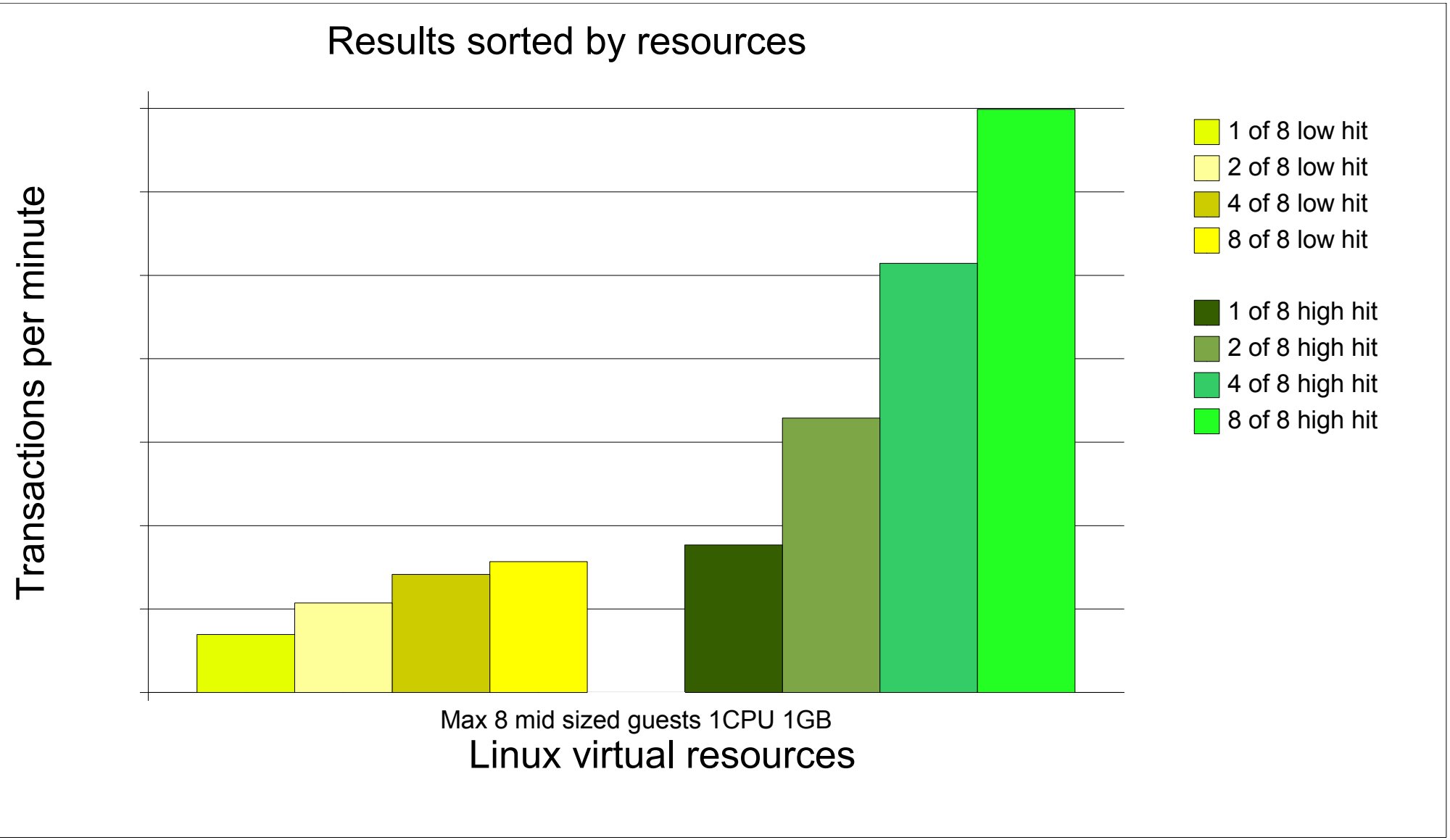
- A few mid sized database servers should perform better than a big single server because they use overall more than 2 GB of memory
- Many small sized servers should not perform worse than the few mid sized servers. Tests with a single small sized server showed notable throughput.



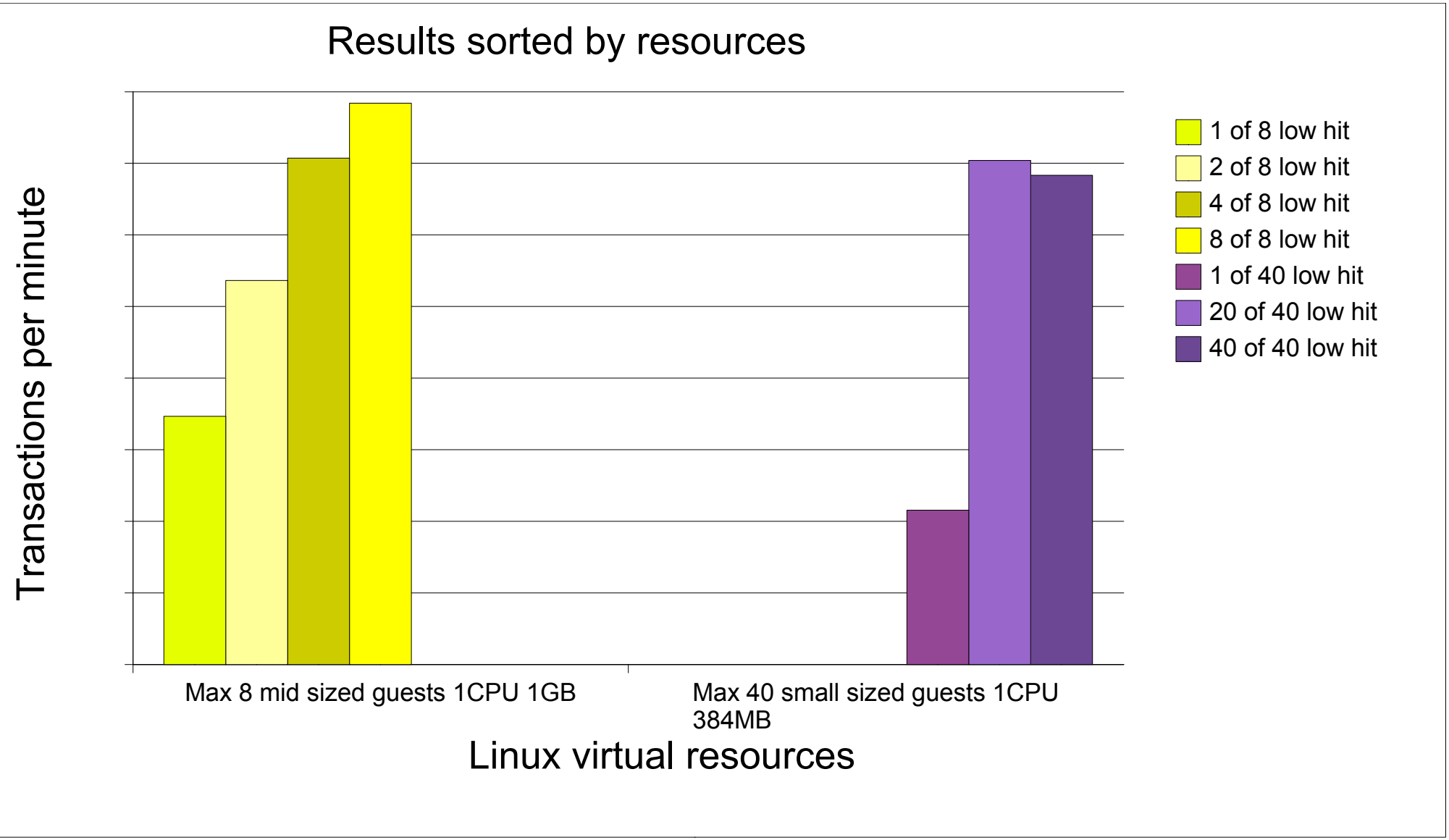
Multi servers test setup

- Few mid sized database servers:
  - 1 virtual CPU, 1 GB memory, 22x 3390-9 disks for database tables
- Many 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

Few mid sized servers results



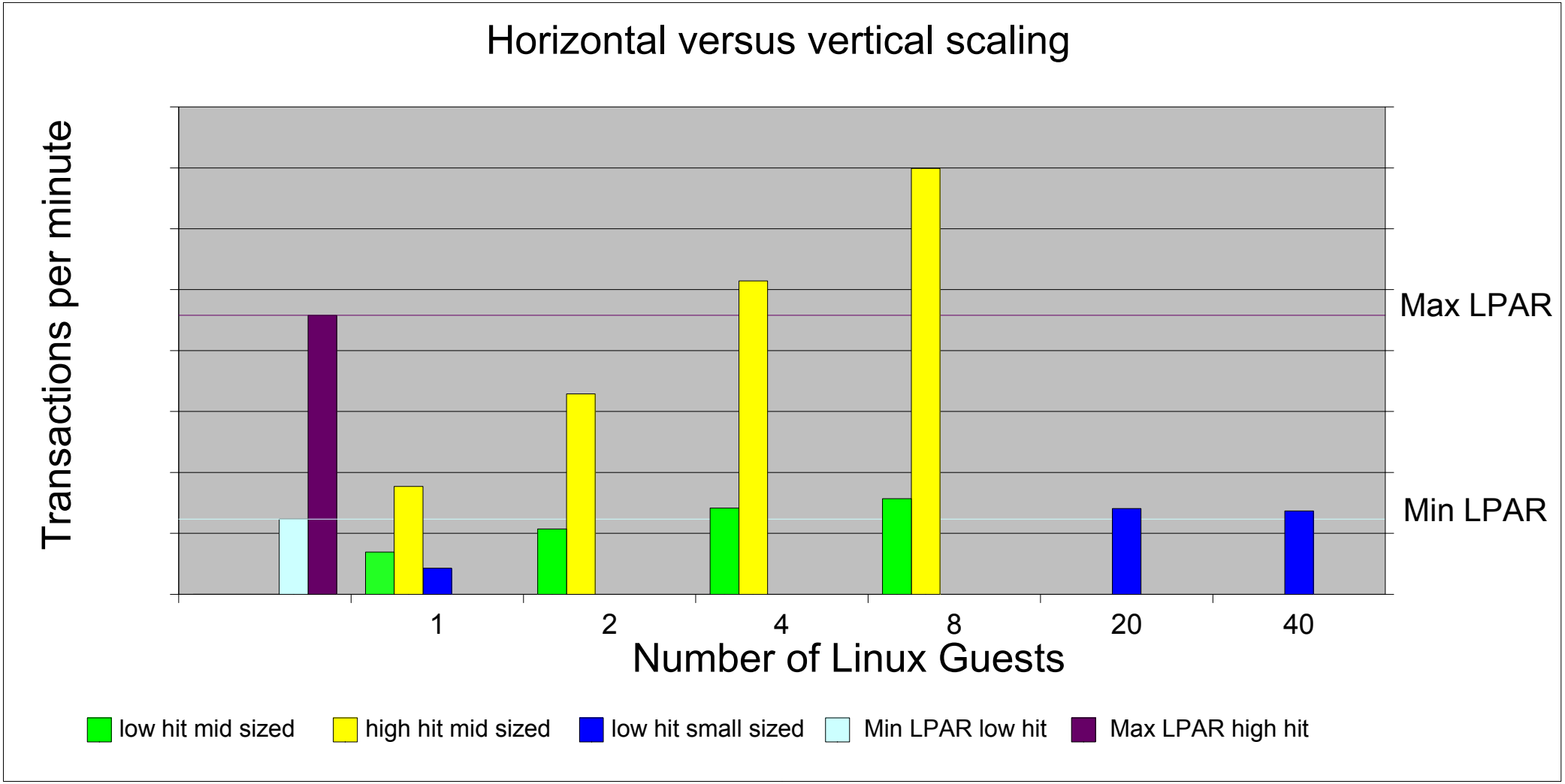
Multi servers results



Multi servers observations

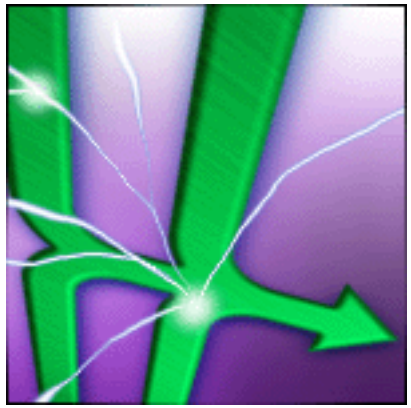
- Total number of disk I/O requests is 8000 SSCH/sec.
  - A busy storage server in a production environment usually runs at 3000 – 5000 SSCH/sec.
  - The test generated almost 2x of usual I/O rates.
- With low hit ratio the performance of many small sized servers and few mid sized servers is equal.

Multi servers versus single server results



Many servers versus single server observations

- High hit ratio
  - The mid sized servers better than one big single server (1.5x)
- Low hit ratio
  - Many small sized database servers perform equally to few mid sized servers and to a single server.



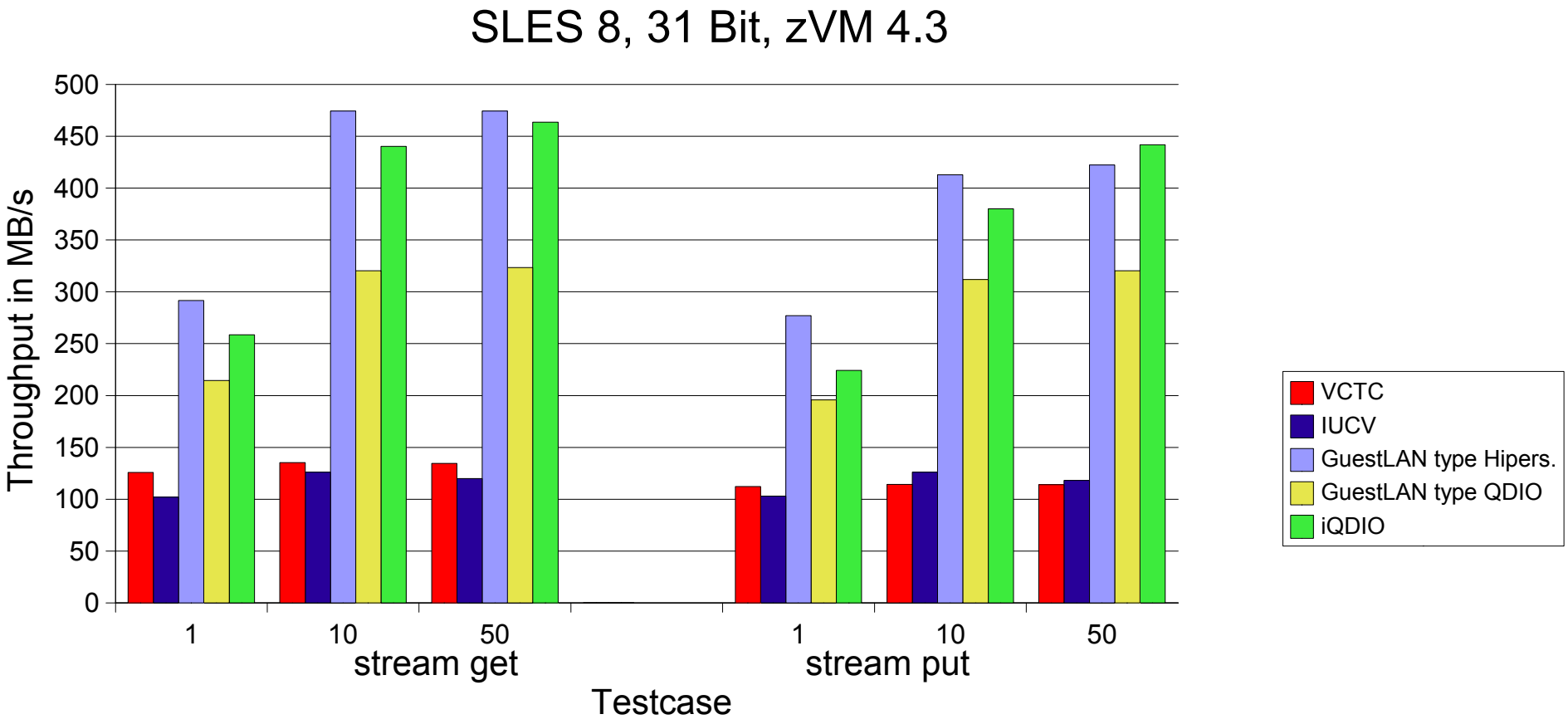
Multi servers performance recommendations

- Remember all recommendations for the single server.
- Provide a big XSTORE in VM (4 GB+).
- For paging provide many entire disks in different ranks as page devices. They should not be used more than 25% on average.
- 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.
- 64bit databases will allow bigger single servers to reach good database buffer hit ratios and reduce high I/O loads.

Networking for your penguin colony



- iQDIO and GuestLAN (GL) type hipersocket show highest throughput
- GL type QDIO a bit worse than GL type hipersocket
- VCTC and IUCV show worst throughput

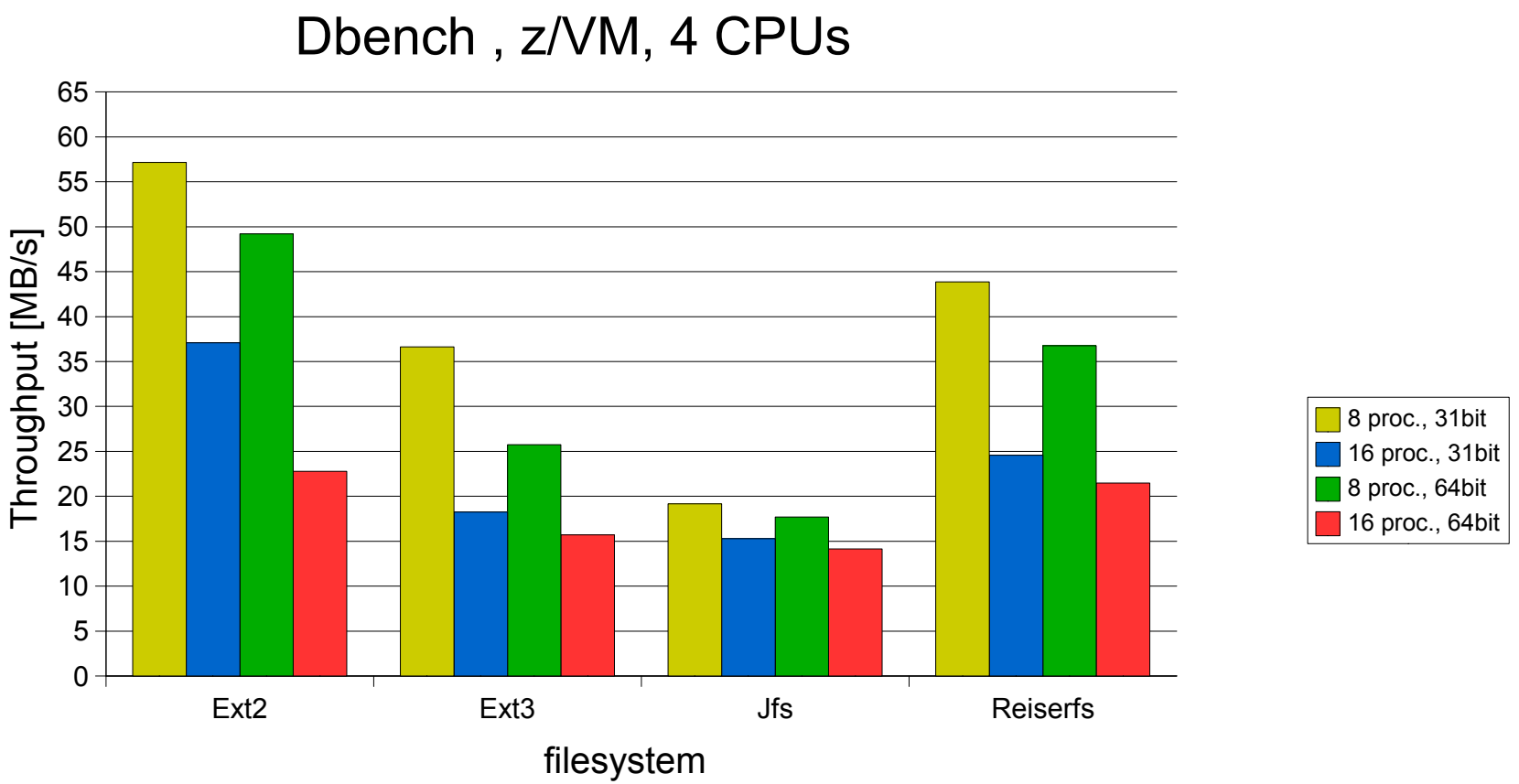
Which network device should I use ?

- Use GuestLAN type hipersocket for inter z/VM guest connections
  - performance comparable to iQDIO
  - easy to use
  - usable on machines older than z800/z900 (zVM 4.3. req.)
  - More connections possible than with iQDIO
- If Multi- and Broadcasts are necessary in your z/VM environment use GuestLAN type QDIO
  - performance a bit less than GuestLAN type hipersocket
  - has packing capability
  - Thin Interrupt will be available with z/VM 4.4

Which network device should I use ? (Cont.)

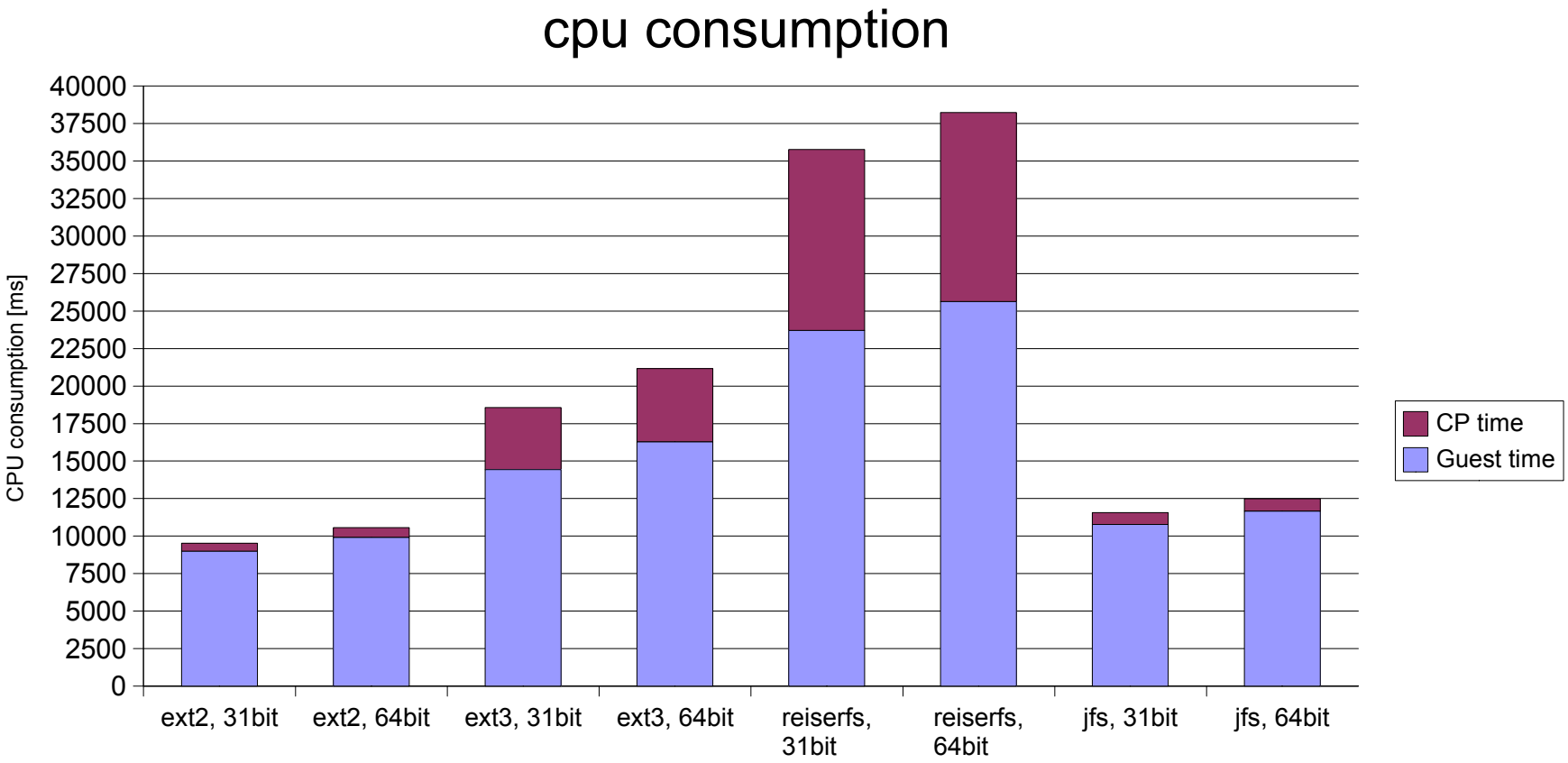
- If your system is low of memory use VCTC or IUCV
  - because each QDIO device (iQDIO, GuestLAN) requires up to 8 MB fixed main memory
- A z/VM guest does not drop from queue Q3 if it uses a QDIO device or CTC device (APAR 63282)
  - apply PTF UM30888 on z/VM 4.3. or UM30889 on z/VM 4.4
- Goal: Find one connection type which fits all topics from above. Can be GuestLAN type QDIO.

Filesystems throughput



- Ext2 shows best results, but provides no journaling
- ReiserFS best throughput of journaling filesystems
- Throughput degradation visible with 64 Bit

Filesystem CPU consumption



- ReiserFS shows highest CPU load, especially CP time is high !
- up to 250000 Diagnose 44 (Voluntary time-slice end) per second
- CPU load higher with Linux 64 bit

Filesystem recommendations for Penguin colony

- Ext 2: no journaling capabilities
  - high I/O rate, long elapse time if many guests do filesystem checks
  - chance of data loss !
- JFS performance sub-optimal
- ReiserFS: high LPAR/CP overhead
- Ext3:
  - good performance
  - low CPU consumption

- Attention: Default during SLES8 installation is ReiserFS !

Visit Us

SHARE
Technology · Connections · Results

- 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