



# RHEL Kernel Performance Optimization, Characterization and Tuning

Larry Woodman

John Shakshober

# Agenda

- **Section 1 – System overview**
- **Section 2 - Analyzing System Performance**
- **Section 3 - Tuning Redhat Enterprise Linux**
- **Section 4 – Performance Analysis and Tuning Examples**
- **References**

# Section 1 - System Overview

- **Processors**
- **NUMA**
- **Memory Management**
- **File System & Disk IO**

# Processors Supported/Tested

- **RHEL4 Limitations**
  - x86 – 16
  - x86\_64 – 8, 512(LargeSMP)
  - ia64 – 8, 64(SGI)
- **RHEL5 Limitations**
  - x86 – 32
  - x86\_64 – 256
  - ia64 - 1024

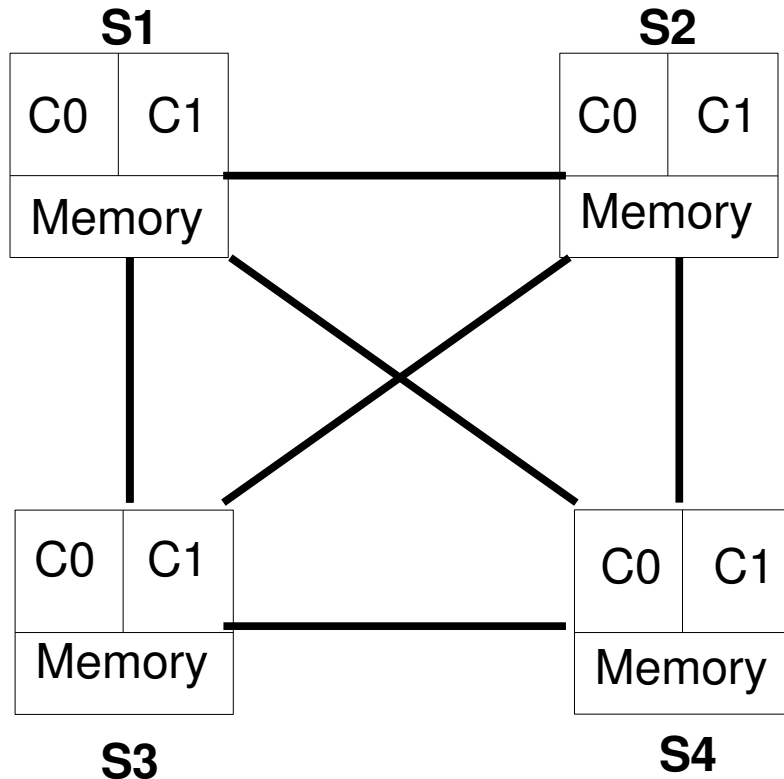
# Processor types

- **Uni-Processor**
- **Symmetric Multi Processor**
- **Multi-Core**
- **Symmetric Multi-Thread**

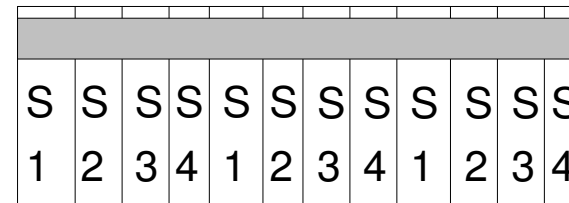
# NUMA Support

- **RHEL3 NUMA Support**
  - **Basic multi-node support**
  - **Local memory allocation**
- **RHEL4 NUMA Support**
  - **NUMA aware memory allocation policy**
  - **NUMA aware memory reclamation**
  - **Multi-core support**
- **RHEL5 NUMA Support**
  - **NUMA aware scheduling**
  - **CPUsets**
  - **NUMA-aware slab allocator**
  - **NUMA-aware hugepages**

# AMD64 System Numa Memory Layout

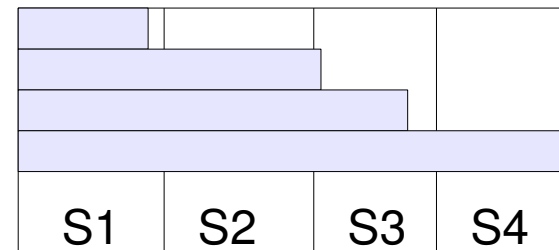


Process on S1C0



interleaved (Non-NUMA)

Process on S1C0



Non-Interleaved (NUMA)

# Memory Management

- **Physical Memory(RAM) Management**
- **Virtual Address Space Maps**
- **Kernel Wired Memory**
- **Reclaimable User Memory**
- **Page Reclaim Dynamics**

# Physical Memory Supported/Tested

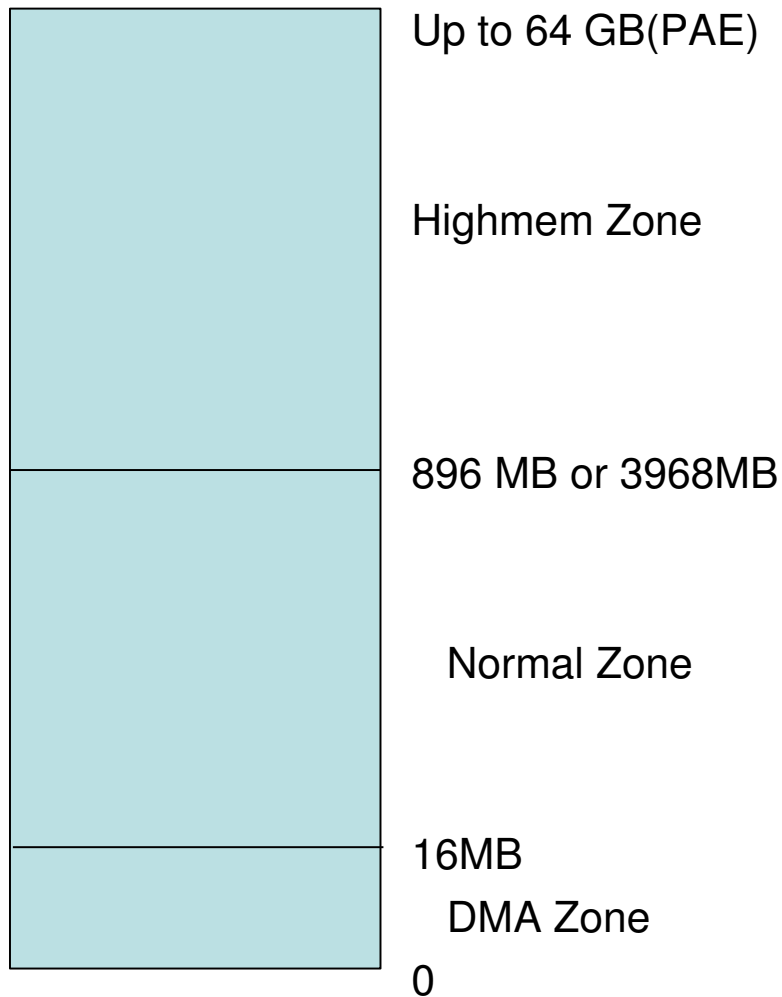
- **RHEL3 Limitations**
  - **x86 – 64GB**
  - **x86\_64 – 64GB**
  - **ia64 – 128GB**
- **RHEL4 Limitations**
  - **x86 – 64GB**
  - **x86\_64 – 128GB**
  - **ia64 – 1TB**
- **RHEL5 Limitations**
  - **x86 – 64GB**
  - **x86\_64 – 256GB**
  - **ia64 - 2TB**

# Physical Memory(RAM) Management

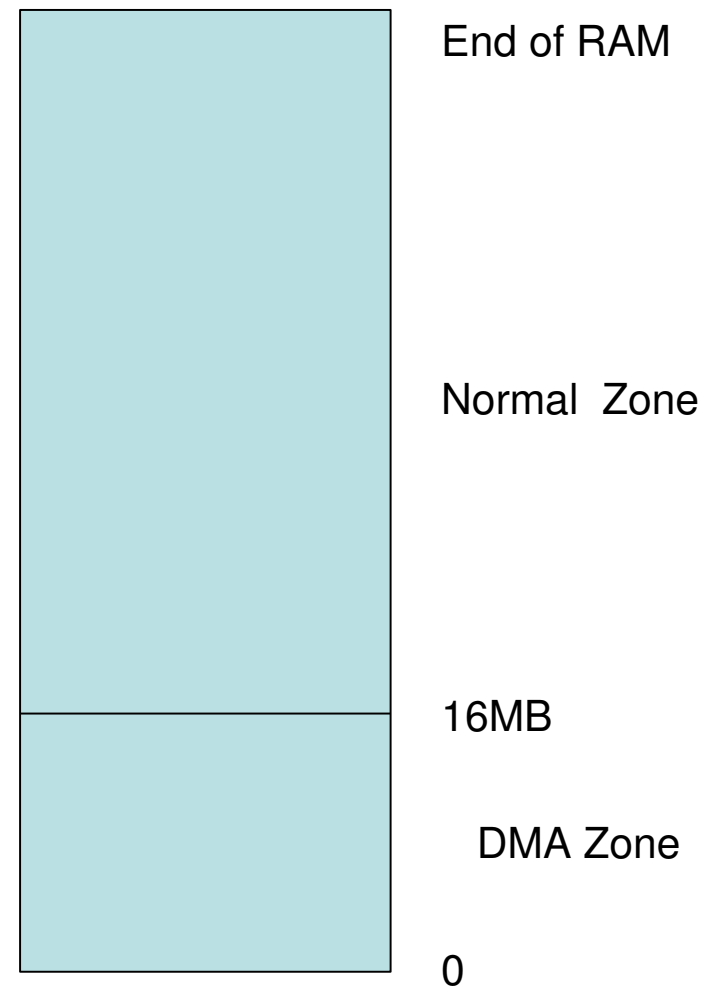
- **Physical Memory Layout**
- **NUMA versus Non-NUMA(UMA)**
- **NUMA Nodes**
  - **Zones**
  - **mem\_map array**
  - **Page lists**
    - **Free list**
    - **Active**
    - **Inactive**

# Memory Zones

## 32-bit



## 64-bit



# Memory Zone Utilization

DMA	Normal	Highmem(x86)
-----	--------	--------------

**24bit I/O**

**Kernel Static**

**User**

**Kernel Dynamic**

**Anonymous**

**slabcache**

**Pagecache**

**bounce buffers**

**Pagetables**

**driver allocations**

**User Overflow**

# Per-Zone Resources

- **RAM**
- **mem\_map**
- **Page lists: free, active and inactive**
- **Page allocation and reclamation**
- **Page reclamation watermarks**

# mem\_map

- Kernel maintains a “page” struct for each 4KB(16KB on IA64 and 64KB for PPC64/RHEL5) page of RAM
- mem\_map is the global array of page structs
- Page struct size:
  - RHEL3 32-bit = 60bytes
  - RHEL3 64-bit = 112bytes
  - RHEL4/RHEL5 32-bit = 32bytes
  - RHEL4/RHEL5 64-bit = 56bytes
- 16GB x86 running RHEL3: ~250MB mem\_map array!!!
- RHEL4 & 5 mem\_map is only about 50% of the RHEL3 mem\_map.

# Per-zone page lists

- **Active List - most recently referenced**
  - **Anonymous-stack, heap, bss**
  - **Pagecache-filesystem data/meta-data**
- **Inactive List - least recently referenced**
  - **Dirty-modified**
  - **Laundry-writeback in progress**
  - **Clean-ready to free**
- **Free**
  - **Coalesced buddy allocator**

# Per zone Free list/buddy allocator lists

- Kernel maintains per-zone free list
- Buddy allocator coalesces free pages into larger physically contiguous pieces

## DMA

1\*4kB 4\*8kB 6\*16kB 4\*32kB 3\*64kB 1\*128kB 1\*256kB 1\*512kB 0\*1024kB 1\*2048kB 2\*4096kB = 11588kB)

## Normal

217\*4kB 207\*8kB 1\*16kB 1\*32kB 0\*64kB 1\*128kB 1\*256kB 1\*512kB 0\*1024kB 0\*2048kB 0\*4096kB = 3468kB)

## HighMem

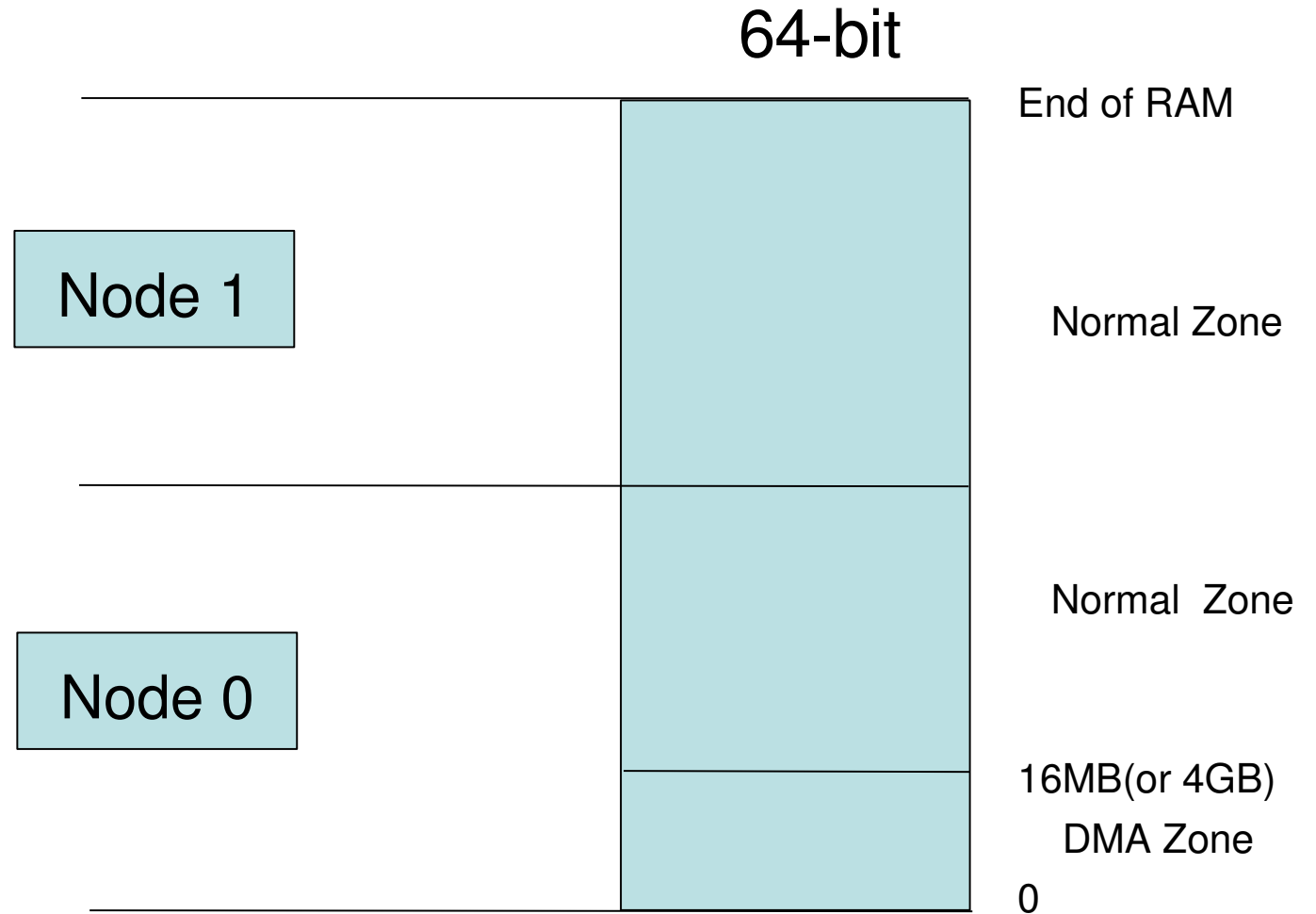
847\*4kB 409\*8kB 17\*16kB 1\*32kB 1\*64kB 1\*128kB 1\*256kB 1\*512kB 0\*1024kB 0\*2048kB 0\*4096kB = 7924kB)

- Memory allocation failures
  - Freelist exhaustion.
  - Freelist fragmentation.

# Per NUMA-Node Resources

- **Memory zones(DMA & Normal zones)**
- **CPUs**
- **IO/DMA capacity**
- **Page reclamation daemon(kswapd#)**

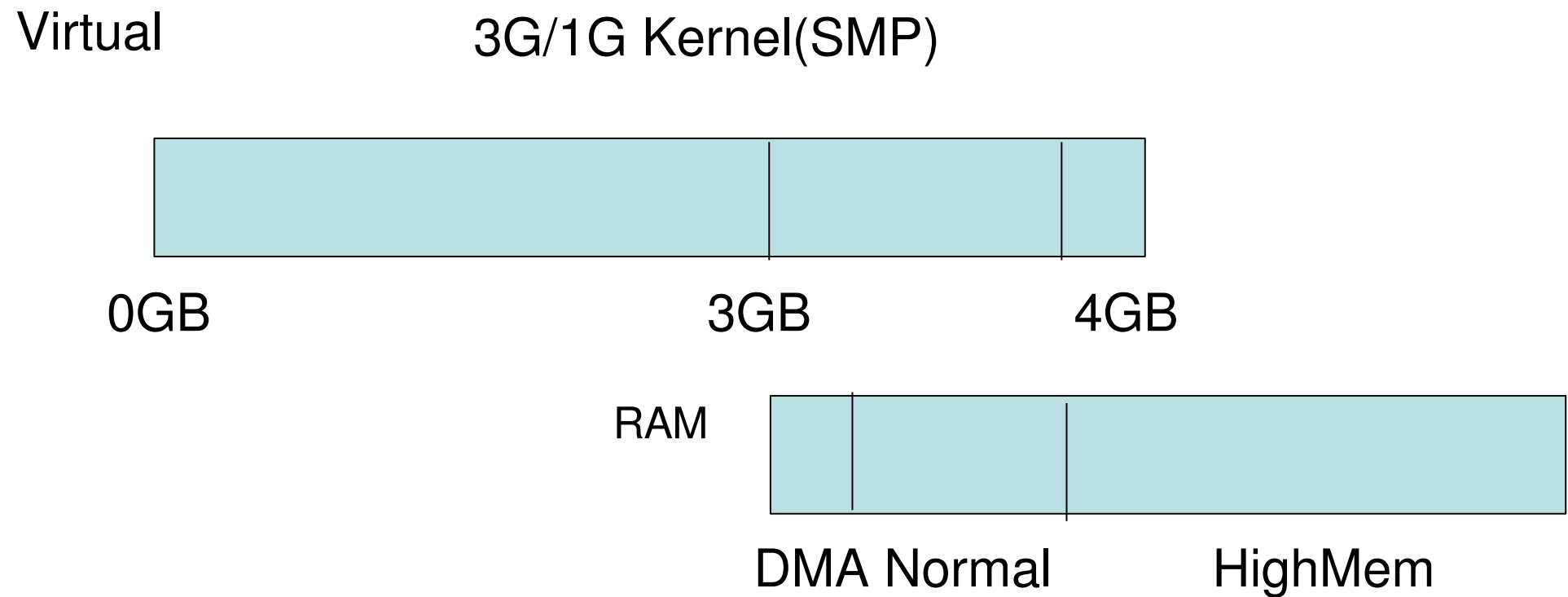
# NUMA Nodes and Zones



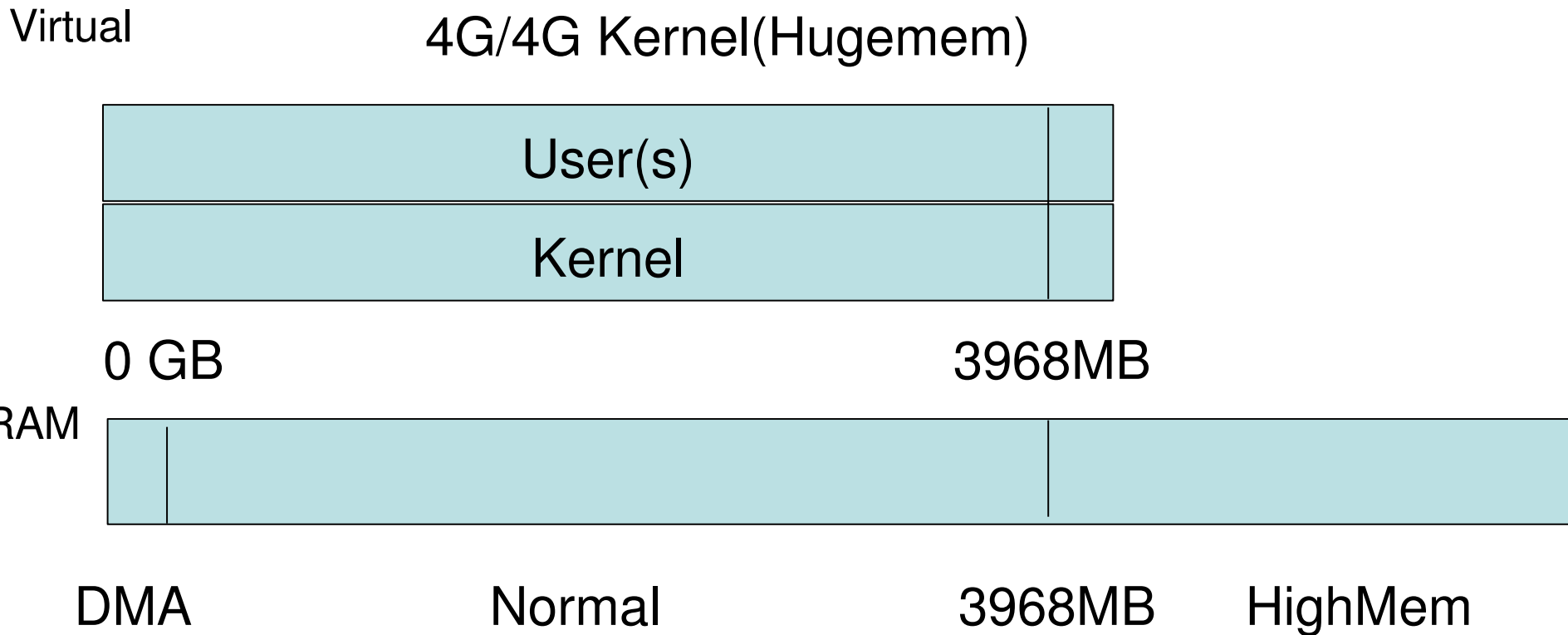
# Virtual Address Space Maps

- **32-bit**
  - **3G/1G address space**
  - **4G/4G address space(RHEL3/4)**
- **64-bit**
  - **X86\_64**
  - **IA64**

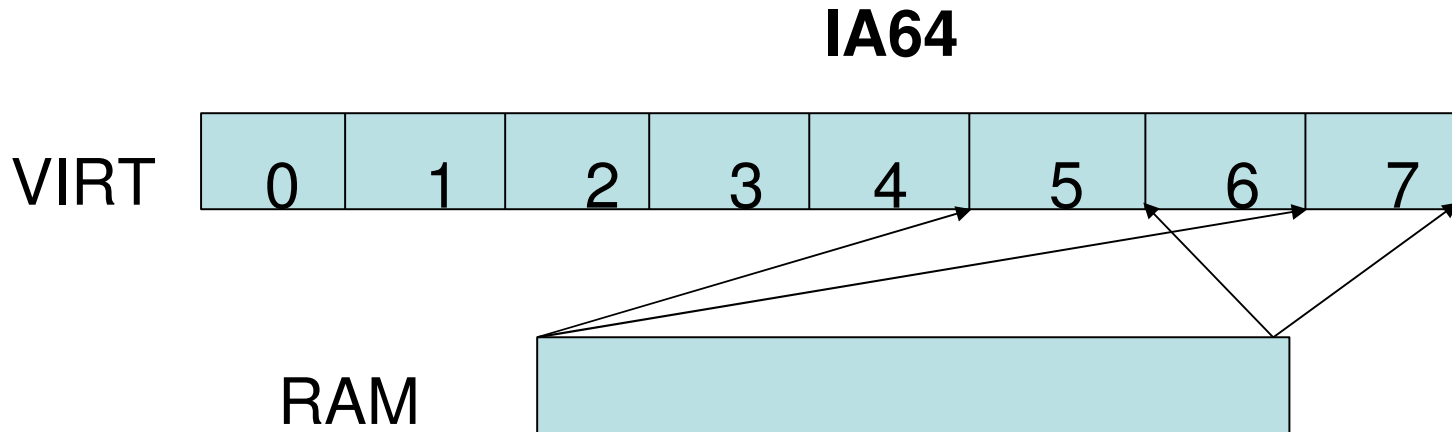
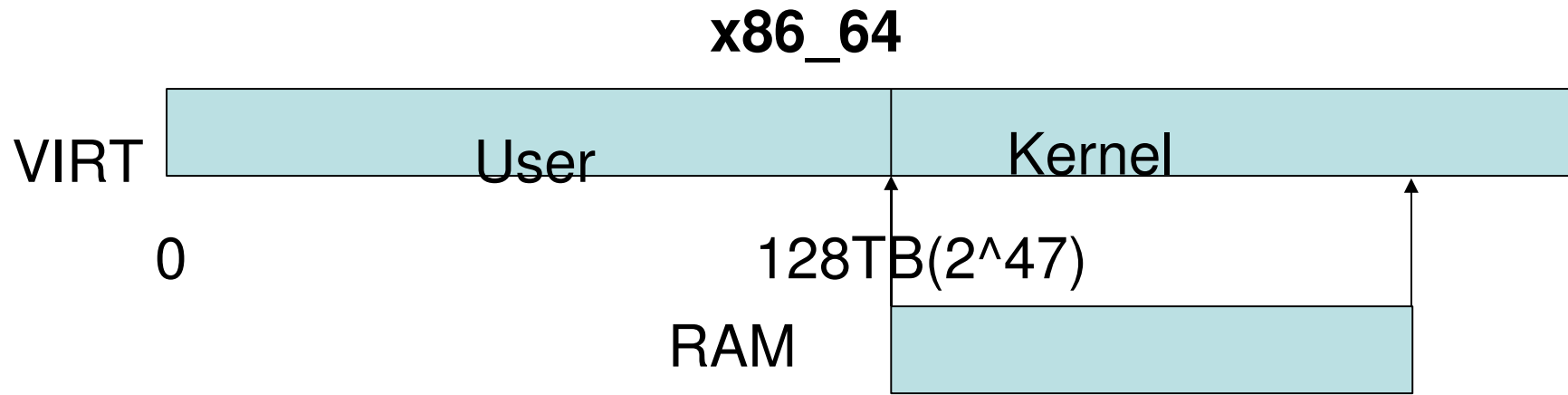
# Linux 32-bit Address Spaces(SMP)



# Linux 32-bit Address Space(Hugemem)



# Linux 64-bit Address Space



# Memory Pressure

32- bit



Kernel Allocations

User Allocations

64- bit



Kernel and User Allocations

# Kernel Memory Pressure

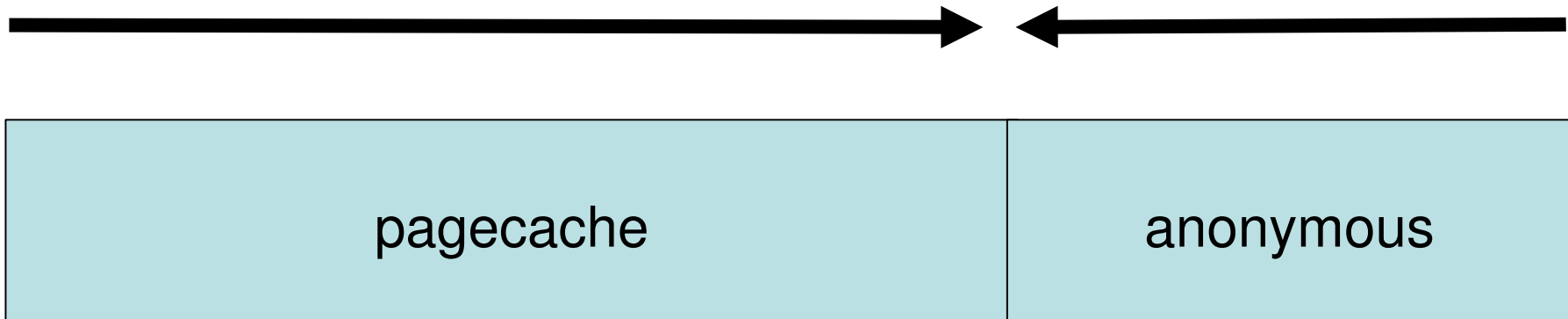
- **Static – Boot-time(DMA and Normal zones)**
  - Kernel text, data, BSS
  - Bootmem allocator, tables and hashes(mem\_map)
- **Dynamic**
  - Slabcache(Normal zone)
    - Kernel data structs
    - Inode cache, dentry cache and buffer header dynamics
  - Pagetables(Highmem/Normal zone)
    - 32bit versus 64bit
- **HughTLBfs(Highmem/Normal zone)**

# User Memory Pressure

## Anonymous/pagecache split

Pagecache Allocations

Page Faults



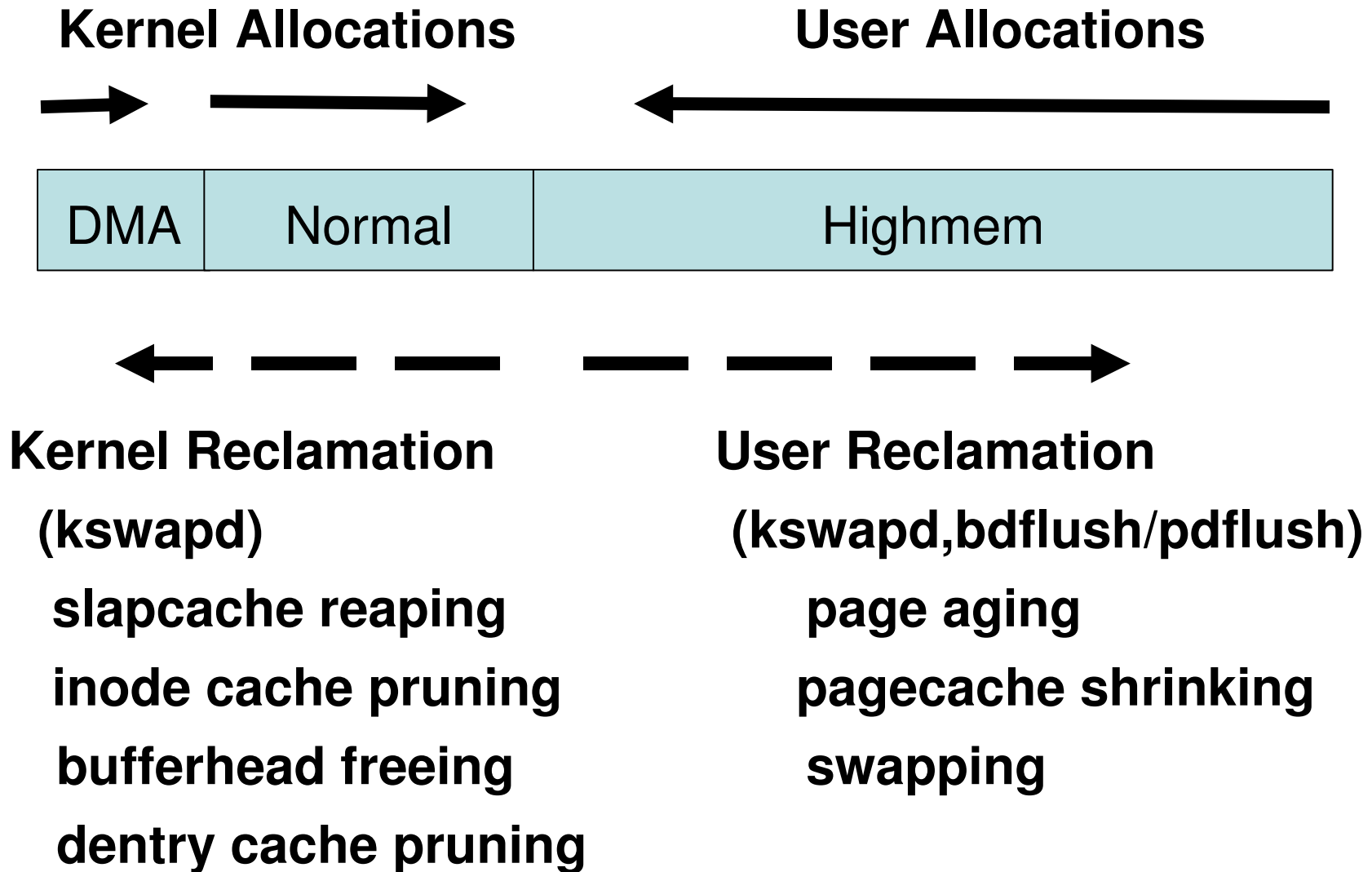
# PageCache/Anonymous memory split

- **Pagecache memory is global and grows when filesystem data is accessed until memory is exhausted.**
- **Pagecache is freed:**
  - **Underlying files are deleted.**
  - **Unmount of the filesystem.**
  - **Kswapd reclaims pagecache pages when memory is exhausted.**
- **Anonymous memory is private and grows on user demand**
  - **Allocation followed by pagefault.**
  - **Swapin.**
- **Anonymous memory is freed:**
  - **Process unmaps anonymous region or exits.**
  - **Kswapd reclaims anonymous pages(swapout) when memory is exhausted**

# PageCache/Anonymous memory split(Cont)

- **Balance between pagecache and anonymous memory.**
  - **Dynamic.**
  - **Controlled via:**
    - **/proc/sys/vm/pagecache.**
    - **/proc/sys/vm/swappiness on RHEL4/RHEL5.**

# 32-bit Memory Reclamation



# 64-bit Memory Reclamation

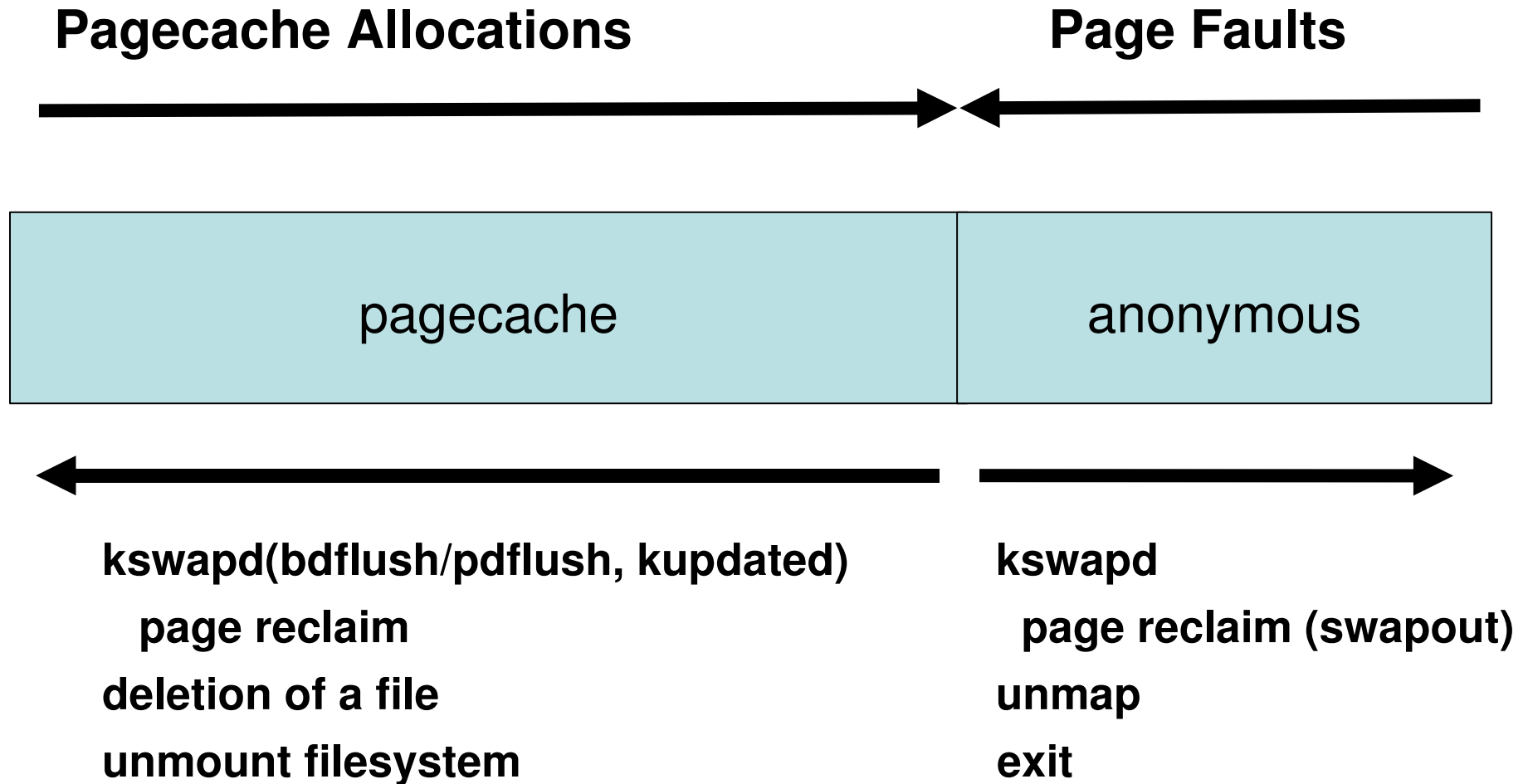


**Kernel and User Allocations**



**Kernel and User Reclamation**

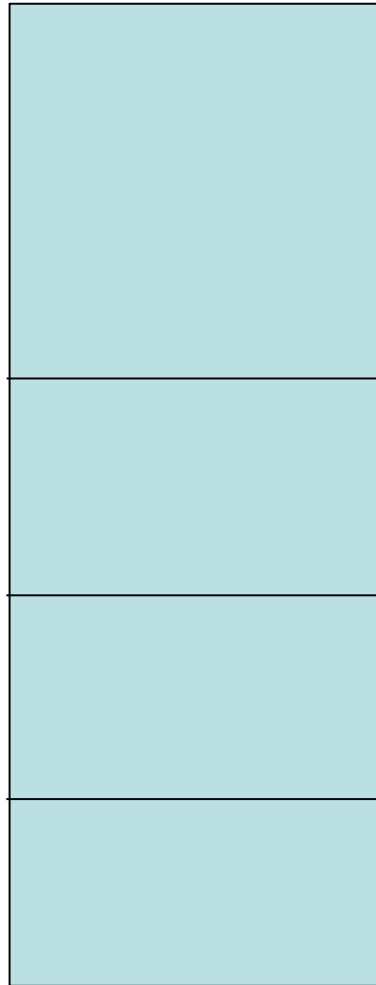
# Anonymous/pagecache reclaiming





# Memory reclaim Watermarks

## Free List



All of RAM

Do nothing

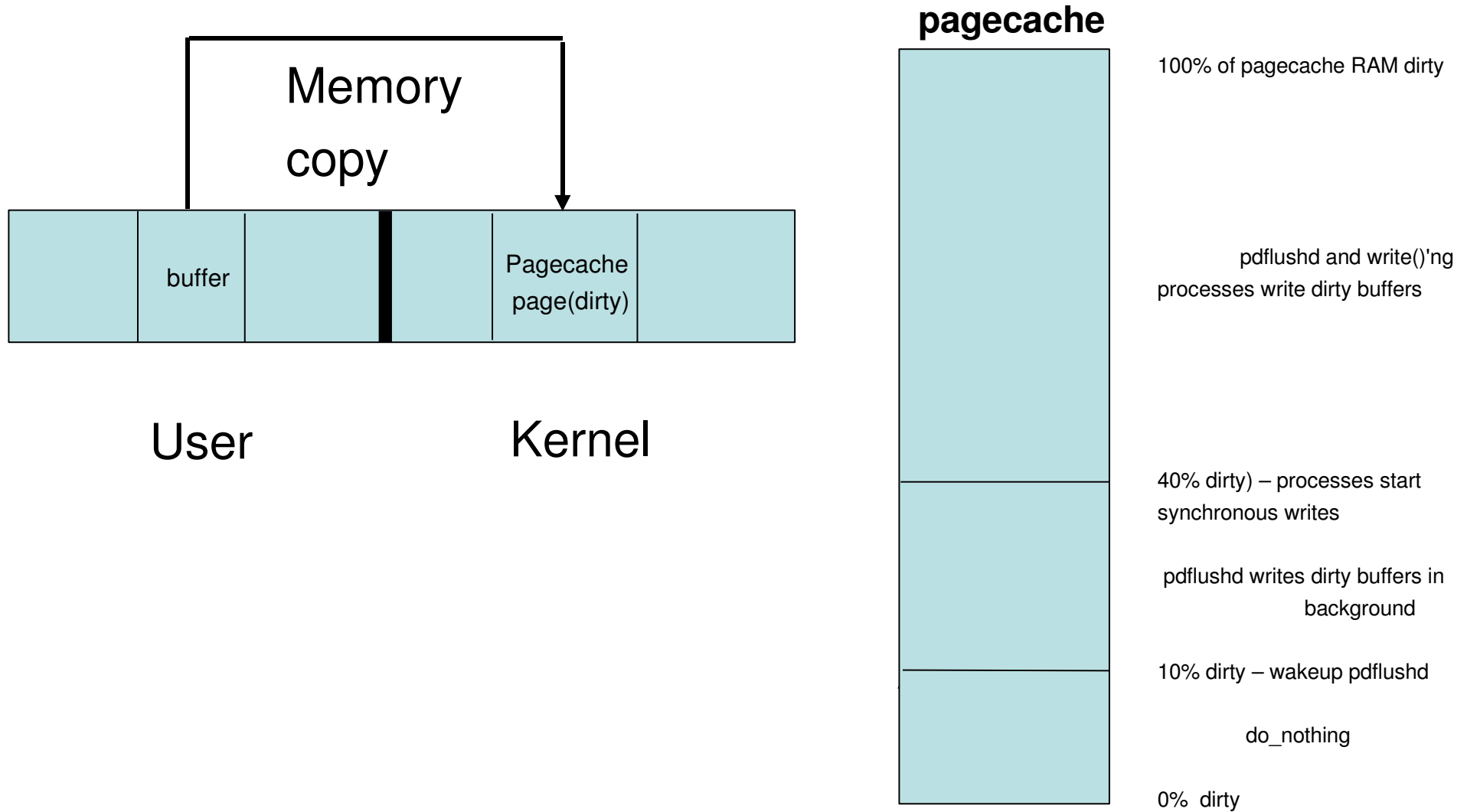
Pages High – kswapd sleeps above High  
kswapd reclaims memory

Pages Low – kswapd wakes up at Low  
kswapd reclaims memory

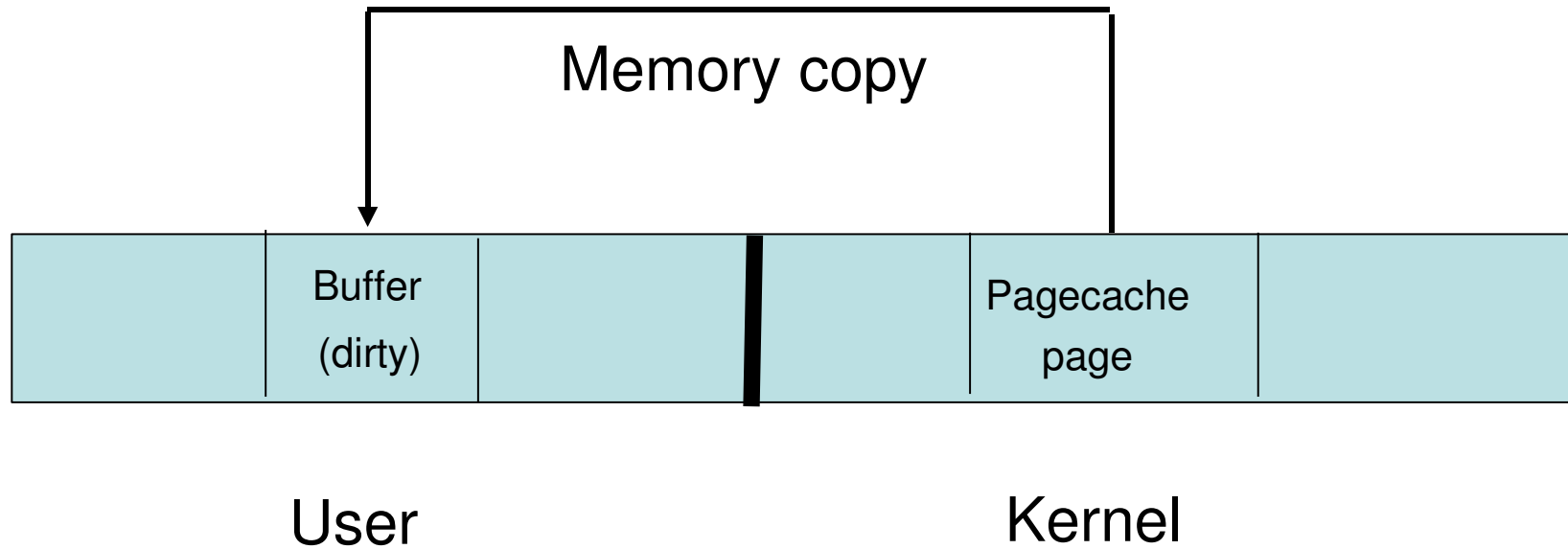
Pages Min – all memory allocators reclaim at Min  
user processes/kswapd reclaim memory

0

# Buffered file system write



# Buffered file system read



# Section 2 - Analyzing System Performance

- **Performance Monitoring Tools**
  - What to run under certain loads
- **Analyzing System Performance**
  - What to look for

# Performance Monitoring Tools

- **Standard Unix OS tools**
  - **Monitoring - cpu, memory, process, disk**
  - **oprofile**
- **Kernel Tools**
  - **/proc, info (cpu, mem, slab), dmesg, AltSysrq**
  - **Profiling - nmi\_watchdog=1, profile=2**
- **Tracing**
  - **strace, ltrace**
  - **dprobe, kprobe**
- **3<sup>rd</sup> party profiling/ capacity monitoring**
  - **Perfmon, Caliper, vtune**
  - **SARcheck, KDE, BEA Patrol, HP Openview**

# Red Hat Top Tools

## ■ CPU Tools

- 1 – top
- 2 – vmstat
- 3 – ps aux
- 4 – mpstat -P all
- 5 – sar -u
- 6 – iostat
- 7 – oprofile
- 8 – gnome-system-monitor
- 9 – KDE-monitor
- 10 – /proc

## ■ Memory Tools

- 1 – top
- 2 – vmstat -s
- 3 – ps aur
- 4 – ipcs
- 5 – sar -r -B -W
- 6 – free
- 7 – oprofile
- 8 – gnome-system-monitor
- 9 – KDE-monitor
- 10 – /proc

## ■ Process Tools

- 1 – top
- 2 – ps -o pmem
- 3 – gprof
- 4 – strace,ltrace
- 5 – sar

## ■ Disk Tools

- 1 – iostat -x
- 2 – vmstat - D
- 3 – sar -DEV #
- 4 – nfsstat
- 5 – NEED MORE!

**top** - press h – help, 1-show cpus, m-memory, t-threads, > -  
column sort

```
top - 09:01:04 up 8 days, 15:22, 2 users, load average: 1.71, 0.39, 0.12
Tasks: 114 total, 1 running, 113 sleeping, 0 stopped, 0 zombie
Cpu0  :  5.3% us,  2.3% sy,  0.0% ni,  0.0% id, 92.0% wa,  0.0% hi,  0.3% si
Cpu1  :  0.3% us,  0.3% sy,  0.0% ni, 89.7% id,  9.7% wa,  0.0% hi,  0.0% si
Mem:   2053860k total, 2036840k used,  17020k free,  99556k buffers
Swap:  2031608k total,   160k used, 2031448k free, 417720k cached
```

PID	USER	PR	NI	VIRT	RES	SHR	S	%CPU	%MEM	TIME+	COMMAND
27830	oracle	16	0	1315m	1.2g	1.2g	D	1.3	60.9	0:00.09	oracle
27802	oracle	16	0	1315m	1.2g	1.2g	D	1.0	61.0	0:00.10	oracle
27811	oracle	16	0	1315m	1.2g	1.2g	D	1.0	60.8	0:00.08	oracle
27827	oracle	16	0	1315m	1.2g	1.2g	D	1.0	61.0	0:00.11	oracle
27805	oracle	17	0	1315m	1.2g	1.2g	D	0.7	61.0	0:00.10	oracle
27828	oracle	15	0	27584	6648	4620	S	0.3	0.3	0:00.17	tpcc.exe
1	root	16	0	4744	580	480	S	0.0	0.0	0:00.50	init
2	root	RT	0	0	0	0	S	0.0	0.0	0:00.11	migration/0
3	root	34	19	0	0	0	S	0.0	0.0	0:00.00	ksoftirqd/0

# vmstat(paging vs swapping)

vmstat 10

procs		memory				swap		io		system			cpu		
r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	wa	id
2	0	0	5483524	200524	234576	0	0	54	63	152	513	0	3	0	96
0	2	0	1697840	200524	3931440	0	0	578	50482	1085	3994	1	22	14	63
3	0	0	7844	200524	5784109	0	0	59330	58946	3243	14430	7	32	18	42

mstat 10

procs		memory				swap		io		system			cpu		
r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	wa	id
2	0	0	5483524	200524	234576	0	0	54	63	152	513	0	3	0	96
0	2	0	1662340	200524	234576	0	0	578	50482	1085	3994	1	22	14	63
3	0	235678	7384	200524	234576	18754	23745	193	58946	3243	14430	7	32	18	42

# Vmstat - IOzone(8GB file with 6GB RAM)

#! deplete memory until pdflush turns on

procs		memory				swap		io		system			cpu		
r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	wa	id
2	0	0	4483524	200524	234576	0	0	54	63	152	513	0	3	0	96
0	2	0	1697840	200524	2931440	0	0	578	50482	1085	3994	1	22	14	63
3	0	0	1537884	200524	3841092	0	0	193	58946	3243	14430	7	32	18	42
0	2	0	528120	200524	6228172	0	0	478	88810	1771	3392	1	32	22	46
0	1	0	46140	200524	6713736	0	0	179	110719	1447	1825	1	30	35	35
2	2	0	50972	200524	6705744	0	0	232	119698	1316	1971	0	25	31	44

....

#! now transition from write to reads

procs		memory				swap		io		system			cpu		
r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	wa	id
1	4	0	51040	200524	6705544	0	0	2	133519	1265	839	0	26	56	18
1	1	0	35064	200524	6712724	0	0	40	118911	1367	2021	0	35	42	23
0	1	0	68264	234372	6647020	0	0	76744	54	2048	4032	0	7	20	73
0	1	0	34468	234372	6678016	0	0	77391	34	1620	2834	0	9	18	72
0	1	0	47320	234372	6690356	0	0	81050	77	1783	2916	0	7	20	73
1	0	0	38756	234372	6698344	0	0	76136	44	2027	3705	1	9	19	72
0	1	0	31472	234372	6706532	0	0	76725	33	1601	2807	0	8	19	73

# iostat -x of same IOzone EXT3 file system

## lostat metrics

rates perf sec

r|w rqm/s – request merged/s

r|w sec/s – 512 byte sectors/s

r|w KB/s – Kilobyte/s

r|w /s – operations/s

sizes and response time

avgrq-sz – average request sz

avequ-sz – average queue sz

await – average wait time ms

svctm – ave service time ms

Linux 2.4.21-27.0.2.ELsmp (node1)

05/09/2005

```
avg-cpu:  %user   %nice    %sys %iowait  %idle
           0.40    0.00    2.63   0.91   96.06
```

```
Device:  rrqm/s wrqm/s  r/s  w/s  rsec/s  wsec/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  svctm   %util
sdi      16164.60   0.00 523.40  0.00 133504.00    0.00 66752.00    0.00  255.07    1.00   1.91   1.88  98.40
sdi      17110.10   0.00 553.90  0.00 141312.00    0.00 70656.00    0.00  255.12    0.99   1.80   1.78  98.40
sdi      16153.50   0.00 522.50  0.00 133408.00    0.00 66704.00    0.00  255.33    0.98   1.88   1.86  97.00
sdi      17561.90   0.00 568.10  0.00 145040.00    0.00 72520.00    0.00  255.31    1.01   1.78   1.76 100.00
```

# SAR

```
[root@localhost redhat]# sar -u 3 3
```

```
Linux 2.4.21-20.EL (localhost.localdomain)
```

```
05/16/2005
```

10:32:28 PM	CPU	%user	%nice	%system	%idle
10:32:31 PM	all	0.00	0.00	0.00	100.00
10:32:34 PM	all	1.33	0.00	0.33	98.33
10:32:37 PM	all	1.34	0.00	0.00	98.66
Average:	all	0.89	0.00	0.11	99.00

```
[root] sar -n DEV
```

```
Linux 2.4.21-20.EL (localhost.localdomain)
```

```
03/16/2005
```

01:10:01 PM	IFACE	rxpck/s	txpck/s	rxbyt/s	txbyt/s	rxcmp/s	txcmp/s
s rxmcst/s							
01:20:00 PM	lo	3.49	3.49	306.16	306.16	0.00	
0.00		0.00					
01:20:00 PM	eth0	3.89	3.53	2395.34	484.70	0.00	
0.00		0.00					
01:20:00 PM	eth1	0.00	0.00	0.00	0.00	0.00	
0.00		0.00					

# free/numastat – memory allocation

```
[root@localhost redhat]# free -l
```

	total	used	free	shared	buffers
cached					
Mem:	511368	342336	169032	0	29712
167408					
Low:	511368	342336	169032	0	0
0					
High:	0	0	0	0	0
0					
-/+ buffers/cache:		145216	366152		
Swap:	1043240	0	1043240		

```
numastat (on 2-cpu x86_64 based system)
```

	node1	node0
numa_hit	9803332	10905630
numa_miss	2049018	1609361
numa_foreign	1609361	2049018
interleave_hit	58689	54749
local_node	9770927	10880901
other_node	2081423	1634090

# ps

```
[root@localhost root]# ps aux
```

```
[root@localhost root]# ps -aux | more
```

USER	PID	%CPU	%MEM	VSZ	RSS	TTY	STAT	START	TIME	COMMAND
root	1	0.1	0.1	1528	516	?	S	23:18	0:04	init
root	2	0.0	0.0	0	0	?	SW	23:18	0:00	[keventd]
root	3	0.0	0.0	0	0	?	SW	23:18	0:00	[kapmd]
root	4	0.0	0.0	0	0	?	SWN	23:18	0:00	[ksoftirqd/0]
root	7	0.0	0.0	0	0	?	SW	23:18	0:00	[bdflush]
root	5	0.0	0.0	0	0	?	SW	23:18	0:00	[kswapd]
root	6	0.0	0.0	0	0	?	SW	23:18	0:00	[kscand]

# pstree

```
init└─/usr/bin/sealer
    └─acpid
    └─atd
    └─auditd└─python
              └─{auditd}
    └─automount──6*[{automount}]
    └─avahi-daemon──avahi-daemon
    └─bonobo-activati──{bonobo-activati}
    └─bt-applet
    └─clock-applet
    └─crond
    └─cupsd──cups-polld
    └─3*[dbus-daemon──{dbus-daemon}]
    └─2*[dbus-launch]
    └─dhclient
```

# mpstat

```
[root@localhost redhat]# mpstat 3 3
```

```
Linux 2.4.21-20.EL (localhost.localdomain) 05/16/2005
```

10:40:34 PM	CPU	%user	%nice	%system	%idle	intr/s
10:40:37 PM	all	3.00	0.00	0.00	97.00	193.67
10:40:40 PM	all	1.33	0.00	0.00	98.67	208.00
10:40:43 PM	all	1.67	0.00	0.00	98.33	196.00
Average:	all	2.00	0.00	0.00	98.00	199.22

# The /proc filesystem

- **/proc**
  - **meminfo**
  - **slabinfo**
  - **cpuinfo**
  - **pid<#>/maps**
  - **vmstat(RHEL4 & RHEL5)**
  - **zoneinfo(RHEL5)**
  - **sysrq-trigger**

# /proc/meminfo(rhel3, 4, 5)

```
RHEL3> cat /proc/meminfo
MemTotal: 509876 kB
MemFree: 17988 kB
MemShared: 0 kB
Buffers: 4728 kB
Cached: 157444 kB
SwapCached: 46576 kB
Active: 222784 kB
ActiveAnon: 118844 kB
ActiveCache: 103940 kB
Inact_dirty: 41088 kB
Inact_laundry: 7640 kB
Inact_clean: 6904 kB
Inact_target: 55680 kB
HighTotal: 0 kB
HighFree: 0 kB
LowTotal: 509876 kB
LowFree: 17988 kB
SwapTotal: 1044184 kB
SwapFree: 945908 kB
CommitLimit: 1299120 kB
Committed_AS: 404920 kB
HugePages_Total: 0
HugePages_Free: 0
Hugepagesize: 2048 kB
```

```
RHEL4> cat /proc/meminfo
MemTotal: 32749568 kB
MemFree: 31313344 kB
Buffers: 29992 kB
Cached: 1250584 kB
SwapCached: 0 kB
Active: 235284 kB
Inactive: 1124168 kB
HighTotal: 0 kB
HighFree: 0 kB
LowTotal: 32749568 kB
LowFree: 31313344 kB
SwapTotal: 4095992 kB
SwapFree: 4095992 kB
Dirty: 0 kB
Writeback: 0 kB
Mapped: 1124080 kB
Slab: 38460 kB
CommitLimit: 20470776 kB
Committed_AS: 1158556 kB
PageTables: 5096 kB
VmallocTotal: 536870911 kB
VmallocUsed: 2984 kB
VmallocChunk: 536867627 kB
HugePages_Total: 0
HugePages_Free: 0
HugePages_Rsvd: 0
Hugepagesize: 2048 kB
```

```
RHEL5> cat /proc/meminfo
MemTotal: 1025220 kB
MemFree: 11048 kB
Buffers: 141944 kB
Cached: 342664 kB
SwapCached: 4 kB
Active: 715304 kB
Inactive: 164780 kB
HighTotal: 0 kB
HighFree: 0 kB
LowTotal: 1025220 kB
LowFree: 11048 kB
SwapTotal: 2031608 kB
SwapFree: 2031472 kB
Dirty: 84 kB
Writeback: 0 kB
AnonPages: 395572 kB
Mapped: 82860 kB
Slab: 92296 kB
PageTables: 23884 kB
NFS_Unstable: 0 kB
Bounce: 0 kB
CommitLimit: 2544216 kB
Committed_AS: 804656 kB
VmallocTotal: 34359738367 kB
VmallocUsed: 263472 kB
VmallocChunk: 34359474711 kB
HugePages_Total: 0
HugePages_Free: 0
HugePages_Rsvd: 0
Hugepagesize: 2048 kB
```

# /proc/slabinfo

slabinfo - version: 2.1

```
# name          <active_objs> <num_objs> <objsize> <objperslab> <pagesperslab> : tunables <limit>
<batchcount> <sharedfactor>: slabdata <active_slabs> <num_slabs> <sharedavail>
nfsd4_delegations      0      0    656      6      1 : tunables    54    27      8 : slabdata      0      0      0
nfsd4_stateids         0      0    128     30      1 : tunables   120    60      8 : slabdata      0      0      0
nfsd4_files            0      0     72     53      1 : tunables   120    60      8 : slabdata      0      0      0
nfsd4_stateowners     0      0    424      9      1 : tunables    54    27      8 : slabdata      0      0      0
nfs_direct_cache      0      0    128     30      1 : tunables   120    60      8 : slabdata      0      0      0
nfs_write_data        36     36    832      9      2 : tunables    54    27      8 : slabdata      4      4      0
nfs_read_data         32     35    768      5      1 : tunables    54    27      8 : slabdata      7      7      0
nfs_inode_cache      1383   1389   1040      3      1 : tunables    24    12      8 : slabdata    463    463      0
nfs_page              0      0    128     30      1 : tunables   120    60      8 : slabdata      0      0      0
fscache_cookie_jar     3      53     72     53      1 : tunables   120    60      8 : slabdata      1      1      0
ip_contrack_expect    0      0    136     28      1 : tunables   120    60      8 : slabdata      0      0      0
ip_contrack           75    130    304     13      1 : tunables    54    27      8 : slabdata     10     10      0
bridge_fdb_cache      0      0     64     59      1 : tunables   120    60      8 : slabdata      0      0      0
rpc_buffers            8      8   2048      2      1 : tunables    24    12      8 : slabdata      4      4      0
rpc_tasks             30     30    384     10      1 : tunables    54    27      8 : slabdata      3      3      0
```

# /proc/cpuinfo

```
[lwoodman]$ cat /proc/cpuinfo
processor      : 0
vendor_id    : GenuineIntel
cpu family   : 6
model        : 15
model name   : Intel(R) Xeon(R) CPU           3060  @ 2.40GHz
stepping     : 6
cpu MHz      : 2394.070
cache size   : 4096 KB
physical id  : 0
siblings     : 2
core id      : 0
cpu cores    : 2
fpu          : yes
fpu_exception : yes
cpuid level  : 10
wp           : yes
flags        : fpu vme de pse tsc msr pae mce cx8 apic sep mtrr pge mca cmov pat pse36 clflush dts
acpi mmx fxsr sse sse2 ss ht tm syscall nx lm constant_tsc pni monitor ds_cpl vmx est tm2 cx16 xtpr
lahf_lm
bogomips     : 4791.41
clflush size : 64
cache_alignment : 64
address sizes : 36 bits physical, 48 bits virtual
power management:
```

# 32-bit /proc/<pid>/maps

```
[root@dhcp83-36 proc]# cat 5808/maps
0022e000-0023b000 r-xp 00000000 03:03 4137068    /lib/tls/libpthread-0.60.so
0023b000-0023c000 rw-p 0000c000 03:03 4137068    /lib/tls/libpthread-0.60.so
0023c000-0023e000 rw-p 00000000 00:00 0
0037f000-00391000 r-xp 00000000 03:03 523285     /lib/libnsl-2.3.2.so
00391000-00392000 rw-p 00011000 03:03 523285     /lib/libnsl-2.3.2.so
00392000-00394000 rw-p 00000000 00:00 0
00c45000-00c5a000 r-xp 00000000 03:03 523268     /lib/ld-2.3.2.so
00c5a000-00c5b000 rw-p 00015000 03:03 523268     /lib/ld-2.3.2.so
00e5c000-00f8e000 r-xp 00000000 03:03 4137064    /lib/tls/libc-2.3.2.so
00f8e000-00f91000 rw-p 00131000 03:03 4137064    /lib/tls/libc-2.3.2.so
00f91000-00f94000 rw-p 00000000 00:00 0
08048000-0804f000 r-xp 00000000 03:03 1046791    /sbin/yppbind
0804f000-08050000 rw-p 00007000 03:03 1046791    /sbin/yppbind
09794000-097b5000 rw-p 00000000 00:00 0
b5fdd000-b5fde000 ---p 00000000 00:00 0
b5fde000-b69de000 rw-p 00001000 00:00 0
b69de000-b69df000 ---p 00000000 00:00 0
b69df000-b73df000 rw-p 00001000 00:00 0
b73df000-b75df000 r--p 00000000 03:03 3270410    /usr/lib/locale/locale-archive
b75df000-b75e1000 rw-p 00000000 00:00 0
bfff6000-c0000000 rw-p ffff8000 00:00 0
```

# 64-bit /proc/<pid>/maps

```
# cat /proc/2345/maps
00400000-0100b000 r-xp 00000000 fd:00 1933328 /usr/sybase/ASE-12_5/bin/dataserver.esd3
0110b000-01433000 rw-p 00c0b000 fd:00 1933328 /usr/sybase/ASE-12_5/bin/dataserver.esd3
01433000-014eb000 rwxp 01433000 00:00 0
40000000-40001000 ---p 40000000 00:00 0
40001000-40a01000 rwxp 40001000 00:00 0
2a95f73000-2a96073000 ---p 0012b000 fd:00 819273 /lib64/tls/libc-2.3.4.so
2a96073000-2a96075000 r--p 0012b000 fd:00 819273 /lib64/tls/libc-2.3.4.so
2a96075000-2a96078000 rw-p 0012d000 fd:00 819273 /lib64/tls/libc-2.3.4.so
2a96078000-2a9607e000 rw-p 2a96078000 00:00 0
2a9607e000-2a98c3e000 rw-s 00000000 00:06 360450 /SYSV0100401e (deleted)
2a98c3e000-2a98c47000 rw-p 2a98c3e000 00:00 0
2a98c47000-2a98c51000 r-xp 00000000 fd:00 819227 /lib64/libnss_files-2.3.4.so
2a98c51000-2a98d51000 ---p 0000a000 fd:00 819227 /lib64/libnss_files-2.3.4.so
2a98d51000-2a98d53000 rw-p 0000a000 fd:00 819227 /lib64/libnss_files-2.3.4.so
2a98d53000-2a98d57000 r-xp 00000000 fd:00 819225 /lib64/libnss_dns-2.3.4.so
2a98d57000-2a98e56000 ---p 00004000 fd:00 819225 /lib64/libnss_dns-2.3.4.so
2a98e56000-2a98e58000 rw-p 00003000 fd:00 819225 /lib64/libnss_dns-2.3.4.so
2a98e58000-2a98e69000 r-xp 00000000 fd:00 819237 /lib64/libresolv-2.3.4.so
2a98e69000-2a98f69000 ---p 00011000 fd:00 819237 /lib64/libresolv-2.3.4.so
2a98f69000-2a98f6b000 rw-p 00011000 fd:00 819237 /lib64/libresolv-2.3.4.so
2a98f6b000-2a98f6d000 rw-p 2a98f6b000 00:00 0
35c7e00000-35c7e08000 r-xp 00000000 fd:00 819469 /lib64/libpam.so.0.77
35c7e08000-35c7f08000 ---p 00008000 fd:00 819469 /lib64/libpam.so.0.77
35c7f08000-35c7f09000 rw-p 00008000 fd:00 819469 /lib64/libpam.so.0.77
35c8000000-35c8011000 r-xp 00000000 fd:00 819468 /lib64/libaudit.so.0.0.0
35c8011000-35c8110000 ---p 00011000 fd:00 819468 /lib64/libaudit.so.0.0.0
35c8110000-35c8118000 rw-p 00010000 fd:00 819468 /lib64/libaudit.so.0.0.0
35c9000000-35c900b000 r-xp 00000000 fd:00 819457 /lib64/libgcc_s-3.4.4-20050721.so.1
35c900b000-35c910a000 ---p 0000b000 fd:00 819457 /lib64/libgcc_s-3.4.4-20050721.so.1
35c910a000-35c910b000 rw-p 0000a000 fd:00 819457 /lib64/libgcc_s-3.4.4-20050721.so.1
7fbffff1000-7fc0000000 rwxp 7fbffff1000 00:00 0
fffffffff600000-ffffffffffffe00000 ---p 00000000 00:00 0
```

# /proc/vmstat(RHEL4/RHEL5)

## cat /proc/vmstat

nr\_anon\_pages 98893  
nr\_mapped 20715  
nr\_file\_pages 120855  
nr\_slab 23060  
nr\_page\_table\_pages 5971  
nr\_dirty 21  
nr\_writeback 0  
nr\_unstable 0  
nr\_bounce 0  
numa\_hit 996729666  
numa\_miss 0  
numa\_foreign 0  
numa\_interleave 87657  
numa\_local 996729666  
numa\_other 0  
pgpgin 2577307  
pgpgout 106131928  
pswpin 0  
pswpout 34  
pgalloc\_dma 198908  
pgalloc\_dma32 997707549  
pgalloc\_normal 0  
pgalloc\_high 0  
pgfree 997909734  
pgactivate 1313196  
pgdeactivate 470908  
pgfault 2971972147  
pgmajfault 8047.

## CONTINUED...

pgrefill\_dma 18338  
pgrefill\_dma32 1353451  
pgrefill\_normal 0  
pgrefill\_high 0  
pgsteal\_dma 0  
pgsteal\_dma32 0  
pgsteal\_normal 0  
pgsteal\_high 0  
pgscan\_kswapd\_dma 7235  
pgscan\_kswapd\_dma32 417984  
pgscan\_kswapd\_normal 0  
pgscan\_kswapd\_high 0  
pgscan\_direct\_dma 12  
pgscan\_direct\_dma32 1984  
pgscan\_direct\_normal 0  
pgscan\_direct\_high 0  
pginodesteal 166  
slabs\_scanned 1072512  
kswapd\_steal 410973  
kswapd\_inodesteal 61305  
pageoutrun 7752  
allocstall 29  
pgrotated 73

# Alt Sysrq M – RHEL3

SysRq : Show Memory

Mem-info:

Zone:DMA freepages: 2929 min: 0 low: 0 high: 0

Zone:Normal freepages: 1941 min: 510 low: 2235 high: 3225

Zone:HighMem freepages: 0 min: 0 low: 0 high: 0

Free pages: 4870 ( 0 HighMem)

( Active: 72404/13523, inactive\_laundry: 2429, inactive\_clean: 1730, free: 4870 )

aa:0 ac:0 id:0 il:0 ic:0 fr:2929

aa:46140 ac:26264 id:13523 il:2429 ic:1730 fr:1941

aa:0 ac:0 id:0 il:0 ic:0 fr:0

1\*4kB 4\*8kB 2\*16kB 2\*32kB 1\*64kB 2\*128kB 2\*256kB 1\*512kB 0\*1024kB 1\*2048kB 2\*4096kB = 11716kB)

1255\*4kB 89\*8kB 5\*16kB 1\*32kB 0\*64kB 1\*128kB 1\*256kB 1\*512kB 1\*1024kB 0\*2048kB 0\*4096kB = 7764kB)

Swap cache: add 958119, delete 918749, find 4611302/5276354, race 0+1

27234 pages of slabcache

244 pages of kernel stacks

1303 lowmem pagetables, 0 highmem pagetables

0 bounce buffer pages, 0 are on the emergency list

Free swap: 598960kB

130933 pages of RAM

0 pages of HIGHMEM

3497 reserved pages

34028 pages shared

39370 pages swap cached

# Alt Sysrq M – RHEL3/NUMA

SysRq : Show Memory

Mem-info:

```
Zone:DMA freepages:      0 min:      0 low:      0 high:      0
Zone:Normal freepages:369423 min:  1022 low:  6909 high:  9980
Zone:HighMem freepages:    0 min:      0 low:      0 high:      0
Zone:DMA freepages:   2557 min:      0 low:      0 high:      0
Zone:Normal freepages:494164 min:  1278 low:  9149 high: 13212
Zone:HighMem freepages:    0 min:      0 low:      0 high:      0
Free pages:      866144 (      0 HighMem)
( Active: 9690/714, inactive_laundry: 764, inactive_clean: 35, free: 866144 )
aa:0 ac:0 id:0 il:0 ic:0 fr:0
aa:746 ac:2811 id:188 il:220 ic:0 fr:369423
aa:0 ac:0 id:0 il:0 ic:0 fr:0
aa:0 ac:0 id:0 il:0 ic:0 fr:2557
aa:1719 ac:4414 id:526 il:544 ic:35 fr:494164
aa:0 ac:0 id:0 il:0 ic:0 fr:0
2497*4kB 1575*8kB 902*16kB 515*32kB 305*64kB 166*128kB 96*256kB 56*512kB 39* 1024kB 30*2048kB 300*4096kB = 1477692kB)
Swap cache: add 288168, delete 285993, find 726/2075, race 0+0
4059 pages of slabcache
146 pages of kernel stacks
388 lowmem pagetables, 638 highmem pagetables
Free swap:      1947848kB
917496 pages of RAM
869386 free pages
30921 reserved pages
21927 pages shared
2175 pages swap cached
Buffer memory:   9752kB
Cache memory:    34192kB
CLEAN: 696 buffers, 2772 kbyte, 51 used (last=696), 0 locked, 0 dirty 0 de lay
DIRTY: 4 buffers, 16 kbyte, 4 used (last=4), 0 locked, 3 dirty 0 delay
```

# Alt Sysrq M – RHEL4 & 5

SysRq : Show Memory

Mem-info:

Free pages: 20128kB (0kB HighMem)

Active:72109 inactive:27657 dirty:1 writeback:0 unstable:0 free:5032 slab:19306 mapped:41755 pagetables:945

DMA free:12640kB min:20kB low:40kB high:60kB active:0kB inactive:0kB present:16384kB pages\_scanned:847  
all\_unreclaimable? yes

protections[]: 0 0 0

Normal free:7488kB min:688kB low:1376kB high:2064kB active:288436kB inactive:110628kB present:507348kB  
pages\_scanned:0 all\_unreclaimable? no

protections[]: 0 0 0

HighMem free:0kB min:128kB low:256kB high:384kB active:0kB inactive:0kB present:0kB pages\_scanned:0  
all\_unreclaimable? no

protections[]: 0 0 0

DMA: 4\*4kB 4\*8kB 3\*16kB 4\*32kB 4\*64kB 1\*128kB 1\*256kB 1\*512kB 1\*1024kB 1\*2048kB 2\*4096kB = 12640kB

Normal: 1052\*4kB 240\*8kB 39\*16kB 3\*32kB 0\*64kB 1\*128kB 0\*256kB 1\*512kB 0\*1024kB 0\*2048kB 0\*4096kB = 7488kB

HighMem: empty

Swap cache: add 52, delete 52, find 3/5, race 0+0

Free swap: 1044056kB

130933 pages of RAM

0 pages of HIGHMEM

2499 reserved pages

71122 pages shared

0 pages swap cached

# Alt Sysrq M – RHEL4 & 5/NUMA

```
Free pages:          16724kB (0kB HighMem)
Active:236461 inactive:254776 dirty:11 writeback:0 unstable:0 free:4181 slab:13679 mapped:34073
pagetables:853
Node 1 DMA free:0kB min:0kB low:0kB high:0kB active:0kB inactive:0kB present:0kB pages_scanned:0
all_unreclaimable? no
protections[]: 0 0 0
Node 1 Normal free:2784kB min:1016kB low:2032kB high:3048kB active:477596kB inactive:508444kB
present:1048548kB pages_scanned:0 all_unreclaimable? no
protections[]: 0 0 0
Node 1 HighMem free:0kB min:128kB low:256kB high:384kB active:0kB inactive:0kB present:0kB pages_scanned:0
all_unreclaimable? no
protections[]: 0 0 0
Node 0 DMA free:11956kB min:12kB low:24kB high:36kB active:0kB inactive:0kB present:16384kB
pages_scanned:1050 all_unreclaimable? yes
protections[]: 0 0 0
Node 0 Normal free:1984kB min:1000kB low:2000kB high:3000kB active:468248kB inactive:510660kB
present:1032188kB pages_scanned:0 all_unreclaimable? no
protections[]: 0 0 0
Node 0 HighMem free:0kB min:128kB low:256kB high:384kB active:0kB inactive:0kB present:0kB pages_scanned:0
all_unreclaimable? no
protections[]: 0 0 0
Node 1 DMA: empty
Node 1 Normal: 0*4kB 0*8kB 30*16kB 10*32kB 1*64kB 1*128kB 1*256kB 1*512kB 1*1024kB 0*2048kB 0*4096kB = 2784kB
Node 1 HighMem: empty
Node 0 DMA: 5*4kB 4*8kB 4*16kB 2*32kB 2*64kB 3*128kB 2*256kB 1*512kB 0*1024kB 1*2048kB 2*4096kB = 11956kB
Node 0 Normal: 0*4kB 0*8kB 0*16kB 0*32kB 1*64kB 1*128kB 1*256kB 1*512kB 1*1024kB 0*2048kB 0*4096kB = 1984kB
Node 0 HighMem: empty
Swap cache: add 44, delete 44, find 0/0, race 0+0
Free swap:          2031432kB
524280 pages of RAM
10951 reserved pages
363446 pages shared
0 pages swap cached
```

# Alt Sysrq T

bash R current 0 1609 1606

(NOTLB)

Call Trace: [

[<c01294b3>] call\_console\_drivers [kernel] 0x63 (0xdb3c5eb4)

[<c01297e3>] printk [kernel] 0x153 (0xdb3c5eec)

[<c01297e3>] printk [kernel] 0x153 (0xdb3c5f00)

[<c010c289>] show\_trace [kernel] 0xd9 (0xdb3c5f0c)

[<c010c289>] show\_trace [kernel] 0xd9 (0xdb3c5f14)

[<c0125992>] show\_state [kernel] 0x62 (0xdb3c5f24)

[<c01cfb1a>] \_\_handle\_sysrq\_nolock [kernel] 0x7a (0xdb3c5f38)

[<c01cfa7d>] handle\_sysrq [kernel] 0x5d (0xdb3c5f58)

[<c0198f43>] write\_sysrq\_trigger [kernel] 0x53 (0xdb3c5f7c)

[<c01645b7>] sys\_write [kernel] 0x97 (0xdb3c5f94)

\* logged in /var/log/messages

# Alt Sysrq W and P

SysRq : Show CPUs

CPU0:

```
ffffff8047ef48 0000000000000000 fffffff80437f10 fffffff8019378b
0000000000000000 0000000000000000 0000000000000000 fffffff801937ba
ffffff8019378b fffffff80022b27 fffffff800551bf 0000000000090000
```

Call Trace:

```
[<ffffff80069572>] show_trace+0x34/0x47
[<ffffff80069675>] _show_stack+0xd9/0xe8
[<ffffff801937ba>] showacpu+0x2f/0x3b
[<ffffff80022b27>] smp_call_function_interrupt+0x57/0x75
[<ffffff8005bf16>] call_function_interrupt+0x66/0x6c
[<ffffff8002fcc2>] unix_poll+0x0/0x96
[<ffffff800551f5>] mwait_idle+0x36/0x4a
[<ffffff80047205>] cpu_idle+0x95/0xb8
[<ffffff8044181f>] start_kernel+0x225/0x22a
[<ffffff8044125b>] _sinittext+0x25b/0x262
```

# oprofile – builtin to RHEL4 & 5 (smp)

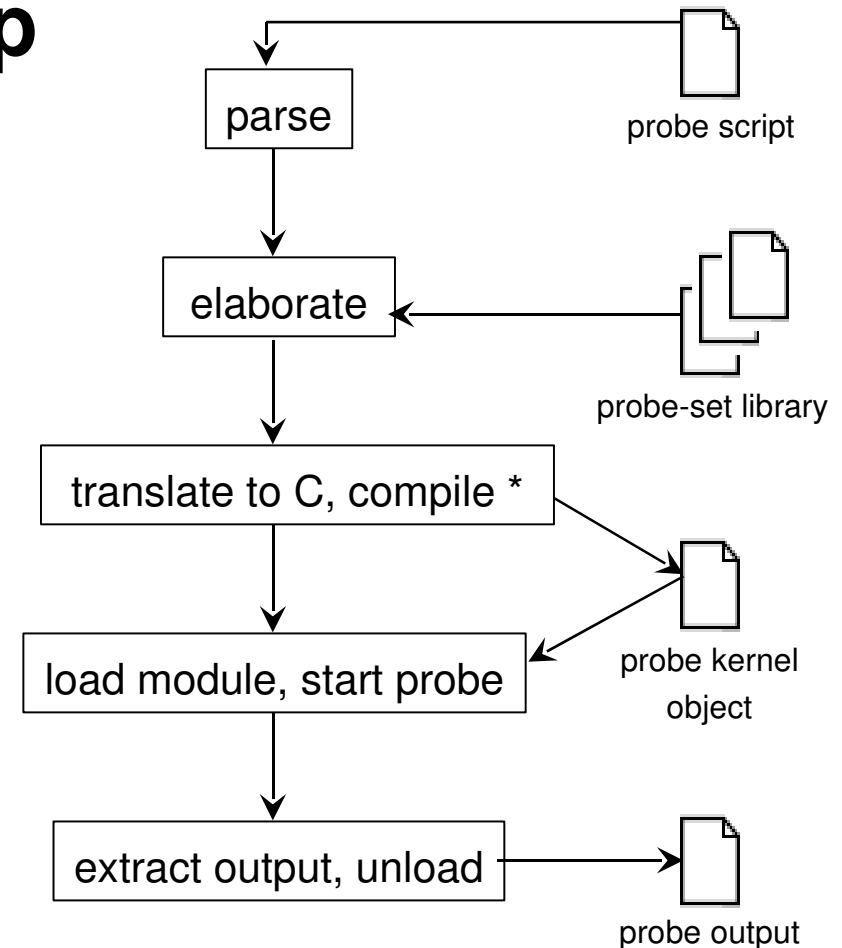
- **opcontrol – on/off data**
  - **--start** start collection
  - **--stop** stop collection
  - **--dump** output to disk
  - **--event=:name:count**
- **Example:**
  - # opcontrol –start**
  - # /bin/time test1 &**
  - # sleep 60**
  - # opcontrol –stop**
  - # opcontrol dump**
- **opreport – analyze profile**
  - **-r** reverse order sort
  - **-t [percentage]** threshold to view
  - **-f /path/filename**
  - **-d** details
- **opannotate**
  - **-s /path/source**
  - **-a /path/assembly**

# oprofile – opcontrol and oprofile cpu\_cycles

```
# CPU: Core 2, speed 2666.72 MHz (estimated)
Counted CPU_CLK_UNHALTED events (Clock cycles when not halted) with a unit mask of 0x00 (Unhalted core c
ycles) count 100000
CPU_CLK_UNHALT...|
  samples|      %|
-----|-----|
397435971 84.6702 vmlinux
 19703064  4.1976 zeus.web
 16914317  3.6034 e1000
 12208514  2.6009 ld-2.5.so
 11711746  2.4951 libc-2.5.so
  5164664  1.1003 sim.cgi
  2333427  0.4971 oprofiled
  1295161  0.2759 oprofile
  1099731  0.2343 zeus.cgi
   968623  0.2064 ext3
   270163  0.0576 jbd
```

# Profiling Tools: SystemTap

- Red Hat, Intel, IBM & Hitachi collaboration
- Linux answer to Solaris Dtrace
- Dynamic instrumentation
- Tool to take a deep look into a running system:
  - Assists in identifying causes of performance problems
  - Simplifies building instrumentation
- Current snapshots available from:  
<http://sources.redhat.com/systemtap>
  - Source for presentations/papers
- Kernel space tracing today, user space tracing under development
- Technology preview status until 5.1



\* Solaris Dtrace is interpretive

# Profiling Tools: SystemTap

- Technology: Kprobes:
  - In current 2.6 kernels
    - Upstream 2.6.12, backported to RHEL4 kernel
  - Kernel instrumentation without recompile/reboot
  - Uses software int and trap handler for instrumentation
- Debug information:
  - Provides map between executable and source code
  - Generated as part of RPM builds
  - Available at: <ftp://ftp.redhat.com>
- Safety: Instrumentation scripting language:
  - No dynamic memory allocation or assembly/C code
  - Types and type conversions limited
  - Restrict access through pointers
- Script compiler checks:
  - Infinite loops and recursion – Invalid variable access

# New Tuning Tools w/ RH MRG

MRG Tuning – using the TUNA – dynamically control  
Device IRQ properties  
CPU affinity / parent and threads  
Scheduling policy

The screenshot displays the Tuna application interface. On the left, a 'Slices' panel shows CPU usage for slices 0 through 7, with values ranging from 11% to 45%. The main window contains a table of IRQs with columns for IRQ, PID, SchedPol, SchedPri, Processor Affinity, and Name. A dialog box titled 'Set IRQ Attributes' is open, showing settings for IRQ 8409 (PID 3464) with a FIFO policy, scheduler priority of 95, and affinity [0-2],[4-7].

IRQ	PID	SchedPol	SchedPri	Processor Affinity	Name
18	615	FIFO	95	[0-2],[4-7]	uhci_hcd:usb3
19	614	FIFO	95	[0-2],[4-7]	uhci_hcd:usb2
8405	3963	FIFO	95	3	eth5
8406	3863	FIFO	95	2	eth4
8407	3763	FIFO	95	1	eth3
8408	3664	FIFO	95	0	eth2
8409	3464	FIFO	95	[0-2],[4-7]	eth1

ID	SchedPol	SchedPri	Proc
4665	Other	0	[0-2]
4694	Other	0	[0-7]
4963	Other	0	[0-2]
4964	Other	0	[0-2]
4965	Other	0	[0-2]
4966	Other	0	[0-2]
4968	Other	0	[0-2],[4-7]
4970	Other	0	[0-2],[4-7]
6150	Other	0	[0-2],[4-7]
6156	Other	0	[0-2],[4-7]
6352	Other	0	[0-2],[4-7]

**Set IRQ Attributes**  
Set attributes for this IRQ:  
IRQ 8409 (PID 3464) Sched FIFO, pri 95, Aff [0-2],[4-7], eth1

Policy	Scheduler priority	Affinity
FIFO	95	[0-2],[4-7]

Cancel OK

# New Tuning Tools w/ RH MRG

MRG Tuning – using the TUNA – dynamically control  
Process affinity / parent and threads  
Scheduling policy

The screenshot displays the Tuna application interface. On the left, a list of processors (0-7) is shown with their respective usage percentages: 0% (0), 0% (1), 0% (2), 0% (3), 39% (4), 45% (5), 53% (6), and 59% (7). The main window shows a table of processes with columns for IRQ, PID, SchedPol, SchedPri, Processor Affinity, and Name. Below this, a detailed view of processes is shown with columns for ID, SchedPol, SchedPri, Processor Affinity, and Command Line. A 'Set Process Attributes' dialog box is open, showing options for 'Set for these processes' (All threads of the selected process), 'Policy' (Other), 'Scheduler priority' (0), and 'Affinity' ([4-7]). The dialog also shows a list of processes with their IDs and command lines.

IRQ	PID	SchedPol	SchedPri	Processor Affinity	Name
18	615	FIFO	95	[0-2],[4-7]	uhci_hcd:usb3
19	614	FIFO	95	[0-2],[4-7]	uhci_hcd:usb2
8405	3963	FIFO	95	3	eth5
8406	3863	FIFO	95	2	eth4
8407	3763	FIFO	95	1	eth3
8408	3664	FIFO	95	0	eth2
8409	3464	FIFO	95	[0-2],[4-7]	eth1

ID	SchedPol	SchedPri	Processor Affinity	Command Line
4665	Other	0	[0-2],[4-7]	/usr/bin/perl -w /usr/sbin/collectl -f /collectl -m -r 00:01,365,144
4694	Other	0	[4-7]	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d
4963	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
4964	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
4965	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
4966	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
4968	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
4970	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
6150	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
6156	Other	0	[0-2],[4-7]	/sbin/mingetty tty1
6352	Other	0	[0-2],[4-7]	/sbin/mingetty tty1

ID	Command Line
4694	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d
4695	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d
4696	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d
4697	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d
4698	/shak/bench/amqp/qpid/cpp/src/.libs/lt-qpid -d

Terminal

root@perf4-1:~ Tuna Set Process Attributes

Applications Places System root Mon Feb 25, 8:28 AM

# Section 3: Tuning

- **How to tune Linux**
- **Capacity tuning**
  - **Fix problems by adding resources**
- **Performance Tuning**
- **Methodology**
  - 1) **Document config**
  - 2) **Baseline results**
  - 3) **While results non-optimal**
    - a) **Monitor/Instrument system/workload**
    - b) **Apply tuning 1 change at a time**
    - c) **Analyze results, exit or loop**
  - 4) **Document final config**

# Tuning - how to set kernel parameters

- `/proc`

```
[root@foobar fs]# cat /proc/sys/kernel/sysrq (see "0")
```

```
[root@foobar fs]# echo 1 > /proc/sys/kernel/sysrq
```

```
[root@foobar fs]# cat /proc/sys/kernel/sysrq (see "1")
```

- `Sysctl` command

```
[root@foobar fs]# sysctl kernel.sysrq
```

```
kernel.sysrq = 0
```

```
[root@foobar fs]# sysctl -w kernel.sysrq=1
```

```
kernel.sysrq = 1
```

```
[root@foobar fs]# sysctl kernel.sysrq
```

```
kernel.sysrq = 1
```

- Edit the `/etc/sysctl.conf` file

```
# Kernel sysctl configuration file for Red Hat Linux
```

```
# Controls the System Request debugging functionality of the kernel
```

```
kernel.sysrq = 1
```

# Capacity Tuning

- **Memory**
  - `/proc/sys/vm/overcommit_memory`
  - `/proc/sys/vm/overcommit_ratio`
  - `/proc/sys/vm/max_map_count`
  - `/proc/sys/vm/nr_hugepages`
- **Kernel**
  - `/proc/sys/kernel/msgmax`
  - `/proc/sys/kernel/msgmnb`
  - `/proc/sys/kernel/msgmni`
  - `/proc/sys/kernel/shmall`
  - `/proc/sys/kernel/shmmax`
  - `/proc/sys/kernel/shmmni`
  - `/proc/sys/kernel/threads-max`
- **Filesystems**
  - `/proc/sys/fs/aio_max_nr`
  - `/proc/sys/fs/file_max`
- **OOM kills**

# OOM kills – swap space exhaustion(RHEL3)

Mem-info:

Zone:DMA freepages: 975 min: 1039 low: 1071 high: 1103

Zone:Normal freepages: 126 min: 255 low: 1950 high: 2925

Zone:HighMem freepages: 0 min: 0 low: 0 high: 0

Free pages: 1101 ( 0 HighMem)

( Active: 118821/401, inactive\_laundry: 0, inactive\_clean: 0, free: 1101 )

aa:1938 ac:18 id:44 il:0 ic:0 fr:974

aa:115717 ac:1148 id:357 il:0 ic:0 fr:126

aa:0 ac:0 id:0 il:0 ic:0 fr:0

6\*4kB 0\*8kB 0\*16kB 1\*32kB 0\*64kB 0\*128kB 1\*256kB 1\*512kB 1\*1024kB 1\*2048kB 0\*4096kB = 3896kB)

0\*4kB 1\*8kB 1\*16kB 1\*32kB 1\*64kB 1\*128kB 1\*256kB 0\*512kB 0\*1024kB 0\*2048kB 0\*4096kB = 504kB)

Swap cache: add 620870, delete 620870, find 762437/910181, race 0+200

2454 pages of slabcache

484 pages of kernel stacks

2008 lowmem pagetables, 0 highmem pagetables

Free swap: 0kB

129008 pages of RAM

0 pages of HIGHMEM

3045 reserved pages

4009 pages shared

0 pages swap cached

# OOM kills – lowmem consumption(RHEL3/x86)

Mem-info:

```
zone:DMA freepages: 2029 min:      0 low:      0 high:      0
Zone:Normal freepages: 1249 min: 1279 low: 4544 high: 6304
Zone:HighMem freepages: 746 min: 255 low: 29184 high: 43776
Free pages:          4024 (    746 HighMem)
( Active: 703448/665000, inactive_laundry: 99878, inactive_clean: 99730, free: 4024 )
aa:0 ac:0 id:0 il:0 ic:0 fr:2029
aa:128 ac:3346 id:113 il:240 ic:0 fr:1249
aa:545577 ac:154397 id:664813 il:99713 ic:99730 fr:746
1*4kB 0*8kB 1*16kB 1*32kB 0*64kB 1*128kB 1*256kB 1*512kB 1*1024kB 1*2048kB 1*4096kB = 8116 kB)
543*4kB 35*8kB 77*16kB 1*32kB 0*64kB 0*128kB 1*256kB 0*512kB 1*1024kB 0*2048kB 0*4096kB = 4996kB)
490*4kB 2*8kB 1*16kB 1*32kB 1*64kB 1*128kB 1*256kB 1*512kB 0*1024kB 0*2048kB 0*4096kB = 29 84kB)
Swap cache: add 4327, delete 4173, find 190/1057, race 0+0
178558 pages of slabcache
1078 pages of kernel stacks
0 lowmem pagetables, 233961 highmem pagetables
Free swap:          8189016kB
2097152 pages of RAM
1801952 pages of HIGHMEM
103982 reserved pages
115582774 pages shared
154 pages swap cached
Out of Memory: Killed process 27100 (oracle).
```

# OOM kills – lowmem consumption(RHEL4&5/x86)

```
Free pages:      9003696kB (8990400kB HighMem)
Active:323264 inactive:346882 dirty:327575 writeback:3686 unstable:0 free:2250924 slab:177094
mapped:15855 pagetables:987
DMA free:12640kB min:16kB low:32kB high:48kB active:0kB inactive:0kB present:16384kB
pages_scanned:149 all_unreclaimable? yes
protections[]: 0 0 0
Normal free:656kB min:928kB low:1856kB high:2784kB active:6976kB inactive:9976kB present:901120kB
pages_scanned:28281 all_unreclaimable? yes
protections[]: 0 0 0
HighMem free:8990400kB min:512kB low:1024kB high:1536kB active:1286080kB inactive:1377552kB
present:12451840kB pages_scanned:0 all_unreclaimable? no
protections[]: 0 0 0
DMA: 4*4kB 4*8kB 3*16kB 4*32kB 4*64kB 1*128kB 1*256kB 1*512kB 1*1024kB 1*2048kB 2*4096kB = 12640kB
Normal: 0*4kB 2*8kB 0*16kB 0*32kB 0*64kB 1*128kB 0*256kB 1*512kB 0*1024kB 0*2048kB 0*4096kB = 656kB
HighMem: 15994*4kB 17663*8kB 11584*16kB 8561*32kB 8193*64kB 1543*128kB 69*256kB 2101*512kB
1328*1024kB 765*2048kB 875*4096kB = 8990400kB
Swap cache: add 0, delete 0, find 0/0, race 0+0
Free swap:      8385912kB
3342336 pages of RAM
2916288 pages of HIGHMEM
224303 reserved pages
666061 pages shared
0 pages swap cached
Out of Memory: Killed process 22248 (httpd).
oom-killer: gfp_mask=0xd0
```

# OOM kills – IO system stall(RHEL4&5/x86)

```
Free pages: 15096kB (1664kB HighMem) Active:34146 inactive:1995536 dirty:255
writeback:314829 unstable:0 free:3774 slab:39266 mapped:31803 pagetables:820
DMA free:12552kB min:16kB low:32kB high:48kB active:0kB inactive:0kB present:16384kB
pages_scanned:2023 all_unreclaimable? yes
protections[]: 0 0 0
Normal free:880kB min:928kB low:1856kB high:2784kB active:744kB inactive:660296kB
present:901120kB pages_scanned:726099 all_unreclaimable? yes
protections[]: 0 0 0
HighMem free:1664kB min:512kB low:1024kB high:1536kB active:135840kB inactive:7321848kB
present:7995388kB pages_scanned:0 all_unreclaimable? no
protections[]: 0 0 0
DMA: 2*4kB 4*8kB 2*16kB 4*32kB 3*64kB 1*128kB 1*256kB 1*512kB 1*1024kB 1*2048kB 2*4096kB =
12552kB
Normal: 0*4kB 18*8kB 14*16kB 0*32kB 0*64kB 0*128kB 0*256kB 1*512kB 0*1024kB 0*2048kB 0*4096kB
= 880kB
HighMem: 6*4kB 9*8kB 66*16kB 0*32kB 0*64kB 0*128kB 0*256kB 1*512kB 0*1024kB 0*2048kB 0*4096kB
= 1664kB
Swap cache: add 856, delete 599, find 341/403, race 0+0
0 bounce buffer pages
Free swap:          4193264kB
2228223 pages of RAM
1867481 pages of HIGHMEM
150341 reserved pages
343042 pages shared
257 pages swap cached
kernel: Out of Memory: Killed process 3450 (hpsmhd).
```

# Eliminating OOMkills

## ■ RHEL3

- `/proc/sys/vm/oom-kill` – number of processes that can be in an OOM kill state at any one time(default 1).

## ■ RHEL4

- `/proc/sys/vm/oom-kill` – oom kill enable/disable flag(default 1).

## ■ RHEL5

- `/proc/<pid>/oom_adj` – per-process OOM adjustment(-17 to +15)
  - Set to -17 to disable that process from being OOM killed
  - Decrease to decrease OOM kill likelihood.
  - Increase to increase OOM kill likelihood.
- `/proc/<pid>/oom_score` – current OOM kill priority.

# General Performance Tuning Considerations

- **Over Committing RAM**
- **Swap device location**
- **Storage device and limits limits**
- **Kernel selection**

# Performance Tuning – (RHEL3)

- **/proc/sys/vm/bdflush**
- **/proc/sys/vm/pagecache**
- **/proc/sys/vm/numa\_memory\_allocator**

# RHEL3 /proc/sys/vm/bdflush

```
int nfract; /* Percentage of buffer cache dirty to activate bdflush */
int ndirty; /* Maximum number of dirty blocks to write out per wake-cycle */
int dummy2; /* old "nrefill" */
int dummy3; /* unused */
int interval; /* jiffies delay between kupdate flushes */
int age_buffer; /* Time for normal buffer to age before we flush it */
int nfract_sync; /* Percentage of buffer cache dirty to activate bdflush synchronously */
int nfract_stop_bdflush; /* Percetange of buffer cache dirty to stop bdflush */
int dummy5; /* unused */
```

## Example:

Settings for Server with ample IO config (default r3 geared for ws)

```
sysctl -w vm.bdflush="50 5000 0 0 200 5000 3000 60 20 0"
```

# RHEL3 /proc/sys/vm/pagecache

- **pagecache.minpercent**
  - Lower limit for pagecache page reclaiming.
  - Kswapd will stop reclaiming pagecache pages below this percent of RAM.
- **pagecache.borrowpercnet**
  - Kswapd attempts to keep the pagecache at this percent of RAM
- **pagecache.maxpercent**
  - Upper limit for pagecache page reclaiming.
  - RHEL2.1 – **hardlimit**, pagecache will not grow above this percent of RAM.
  - RHEL3 – **kswapd** only reclaims pagecache pages above this percent of RAM.
  - Increasing **maxpercent** will increase swapping

Example: `echo "1 10 50" > /proc/sys/vm/pagecache`

# RHEL3 /proc/sys/vm/numa\_memory\_allocator

>numa=on (default)

```
-----  
Zone:Normal freepages: 10539 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages: 10178 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages: 10445 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages:856165 min: 1279 low: 17342 high: 25501  
Swap cache: add 2633120, delete 2553093, find 1375365/1891330, race 0+0  
-----
```

>numa=off

```
-----  
Zone:Normal freepages:861136 min: 1279 low: 30950 high: 63065  
Swap cache: add 0, delete 0 find 0/0, race 0+0  
-----
```

>numa=on and /proc/sys/vm/numa\_memory\_allocator set to 1

```
-----  
Zone:Normal freepages: 17406 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages: 17406 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages: 17406 min: 1279 low: 17406 high: 25597  
Zone:Normal freepages:85739 min: 1279 low: 17342 high: 25501  
Swap cache: add 0, delete 0 find 0/0, race 0+0  
-----
```

# Performance Tuning – (RHEL4 and RHEL5)

- **/proc/sys/vm/swappiness**
- **/proc/sys/vm/min\_free\_kbytes**
- **/proc/sys/vm/dirty\_ratio**
- **/proc/sys/vm/dirty\_background\_ratio**
- **/proc/sys/vm/pagecache**

# RHEL4 /proc/sys/vm/swappiness

- Controls how aggressively the system reclaims “mapped” memory:
  - Anonymous memory - swapping
  - Mapped file pages – writing if dirty and freeing
  - System V shared memory - swapping
- Decreasing: more aggressive reclaiming of unmapped pagecache memory
- Increasing: more aggressive swapping of mapped memory

Sybase server with /proc/sys/vm/swappiness set to 60(default)

```
procs -----memory----- ---swap--  -----io----  --system--  -----cpu-----
r  b   swpd   free   buff  cache   si   so    bi   bo    in    cs   us  sy  id  wa
5  1  643644 26788   3544 32341788 880 120   4044 7496  1302 20846 25 34 25 16
```

Sybase server with /proc/sys/vm/swappiness set to 10

```
procs -----memory----- ---swap--  -----io----  --system--  -----cpu-----
r  b   swpd   free   buff  cache   si   so    bi   bo    in    cs   us  sy  id  wa
8  3     0   24228   6724 32280696  0    0   23888 63776  1286 20020 24 38 13 26
```

# RHEL4&5 /proc/sys/vm/min\_free\_kbytes

- Directly controls the page reclaim watermarks in KB

```
# echo 1024 > /proc/sys/vm/min_free_kbytes
```

-----

```
Node 0 DMA free:4420kB min:8kB low:8kB high:12kB
```

```
Node 0 DMA32 free:14456kB min:1012kB low:1264kB high:1516kB
```

-----

```
echo 2048 > /proc/sys/vm/min_free_kbytes
```

-----

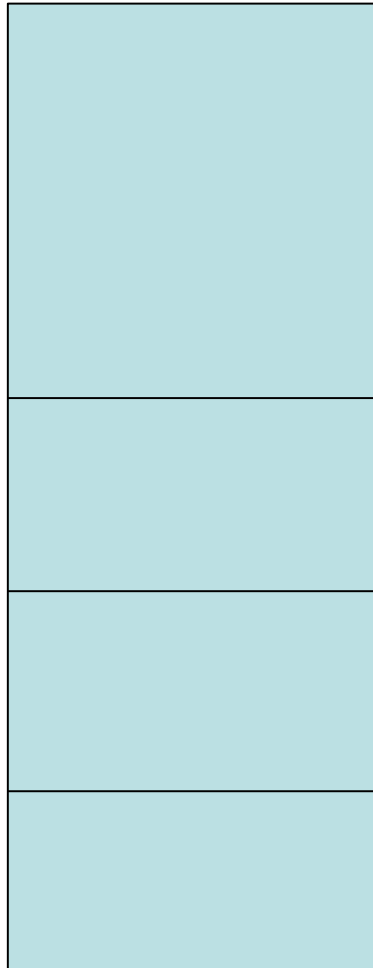
```
Node 0 DMA free:4420kB min:20kB low:24kB high:28kB
```

```
Node 0 DMA32 free:14456kB min:2024kB low:2528kB high:3036kB
```

-----

# Memory reclaim Watermarks - min\_free\_kbytes

## Free List



All of RAM

Do nothing

Pages High – kswapd sleeps above High  
kswapd reclaims memory

Pages Low – kswapd wakes up at Low  
kswapd reclaims memory

Pages Min – all memory allocators reclaim at Min  
user processes/kswapd reclaim memory

0

# RHEL4&5 /proc/sys/vm/dirty\_ratio

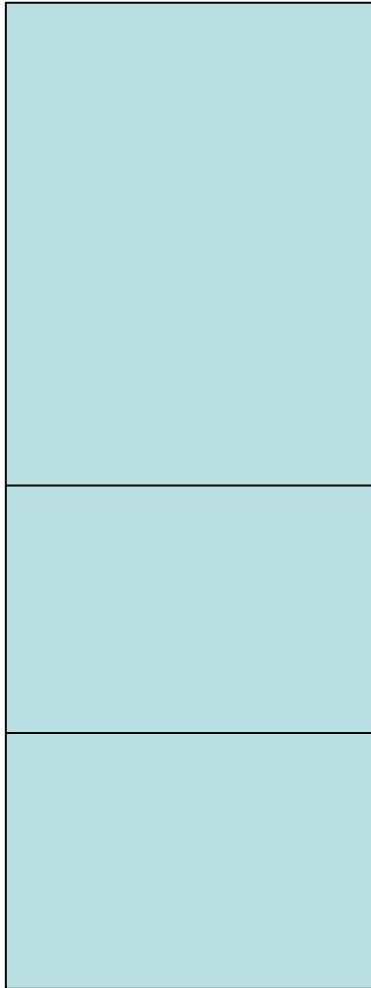
- **Absolute limit to percentage of dirty pagecache memory**
  - **Default is 40%**
  - **Lower means less dirty pagecache and smaller IO streams**
  - **Higher means more dirty pagecache and larger IO streams**

# RHEL4&5 /proc/sys/vm/dirty\_background\_ratio

- Controls when dirty pagecache memory starts getting written.
  - Default is 10%
  - Lower
    - pdflush starts earlier
    - less dirty pagecache and smaller IO streams
  - Higher
    - pdflush starts later
    - more dirty pagecache and larger IO streams

# dirty\_ratio and dirty\_background\_ratio

## pagecache



100% of pagecache RAM dirty

pdflushd and write()'ng processes write dirty buffers

dirty\_ratio(40% of RAM dirty) – processes start synchronous writes

pdflushd writes dirty buffers in background

dirty\_background\_ratio(10% of RAM dirty) – wakeup pdflushd

do\_nothing

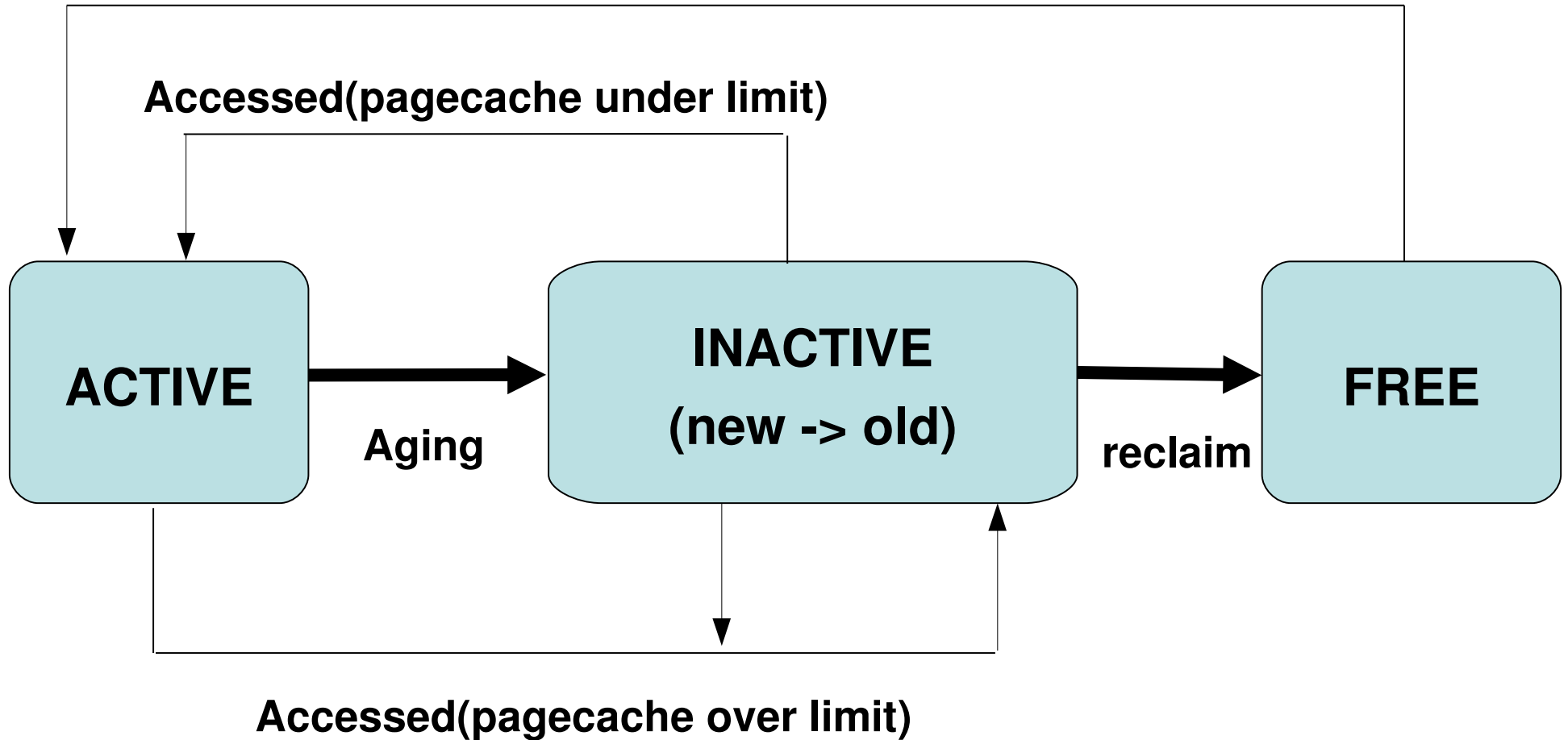
0% of pagecache RAM dirty

# RHEL4&5 /proc/sys/vm/pagecache

- Controls when pagecache memory is deactivated.
- Default is 100%
- Lower
  - Prevents swapping out anonymous memory
- Higher
  - Favors pagecache pages
  - Disabled at 100%

# Pagecache Tuning

## Filesystem/pagecache Allocation



# (Hint)flushing the pagecache

- [tmp]# echo 1 > /proc/sys/vm/drop\_caches

```
procs -----memory----- ---swap-- -----io----- --system-- -----cpu-----
r  b   swpd   free   buff  cache   si   so    bi    bo    in    cs  us  sy  id  wa
0  0     224   57184 107808 3350196    0    0     0    56  1136   212  0  0  83  17
0  0     224   57184 107808 3350196    0    0     0     0  1039   198  0  0 100  0
0  0     224   57184 107808 3350196    0    0     0     0  1021   188  0  0 100  0
0  0     224   57184 107808 3350196    0    0     0     0  1035   204  0  0 100  0
0  0     224   57248 107808 3350196    0    0     0     0  1008   164  0  0 100  0
3  0     224 2128160    176 1438636    0    0     0     0  1030   197  0  0  15  85  0
0  0     224 3610656    204  34408    0    0    28    36  1027   177  0  0  32  67  2
0  0     224 3610656    204  34408    0    0     0     0  1026   180  0  0 100  0
0  0     224 3610720    212  34400    0    0     8     0  1010   183  0  0  99  1
```

# (Hint)flushing the slabcache

- [tmp]# echo 2 > /proc/sys/vm/drop\_caches

```
[tmp]# cat /proc/meminfo
```

```
MemTotal:    3907444 kB
```

```
MemFree:     3604576 kB
```

```
Slab:        115420 kB
```

```
Hugepagesize: 2048 kB
```

```
tmp]# cat /proc/meminfo
```

```
MemTotal:    3907444 kB
```

```
MemFree:     3604576 kB
```

```
Slab:        115420 kB
```

```
Hugepagesize: 2048 kB
```

# RHEL3 kernel selection

- **x86**
  - **Standard kernel(no PAE, 3G/1G)**
    - **UP systems with  $\leq$  4GB RAM**
    - **PAE costs  $\sim$ 5% in performance**
  - **SMP kernel(PAE, 3G/1G)**
    - **SMP systems with  $<$   $\sim$ 12GB RAM**
    - **Highmem/Lowmem ratio  $\leq$  10:1**
    - **4G/4G costs  $\sim$ 5%**
  - **Hugemem kernel(PAE, 4G/4G)**
    - **SMP systems  $>$   $\sim$ 12GB RAM**
- **X86\_64**
  - **Standard kernel for UP systems**
  - **SMP kernel for SMP systems**

# RHEL4 kernel selection

- **x86**
  - **Standard kernel(no PAE, 3G/1G)**
    - **UP systems with  $\leq$  4GB RAM**
  - **SMP kernel(PAE, 3G/1G)**
    - **SMP systems with  $<$  ~16GB RAM**
    - **Highmem/Lowmem ratio  $\leq$  16:1**
  - **Hugemem kernel(PAE, 4G/4G)**
    - **SMP systems  $>$  ~16GB RAM**
- **X86\_64**
  - **Standard kernel for UP systems**
  - **SMP kernel for systems with up to 8 CPUs**
  - **LargeSMP kernel for systems up to 512 CPUs**

# RHEL5 kernel selection

- **x86**
  - **Standard kernel(no PAE, 3G/1G)**
    - **UP and SMP systems with  $\leq$  4GB RAM**
  - **PAE kernel(PAE, 3G/1G)**
    - **UP and SMP systems with  $>$ 4GB RAM**
- **X86\_64**
  - **Standard kernel for all systems**
- **IA64**
  - **Standard kernel for all systems**

# Problem - 16GB x86 running SMP kernel

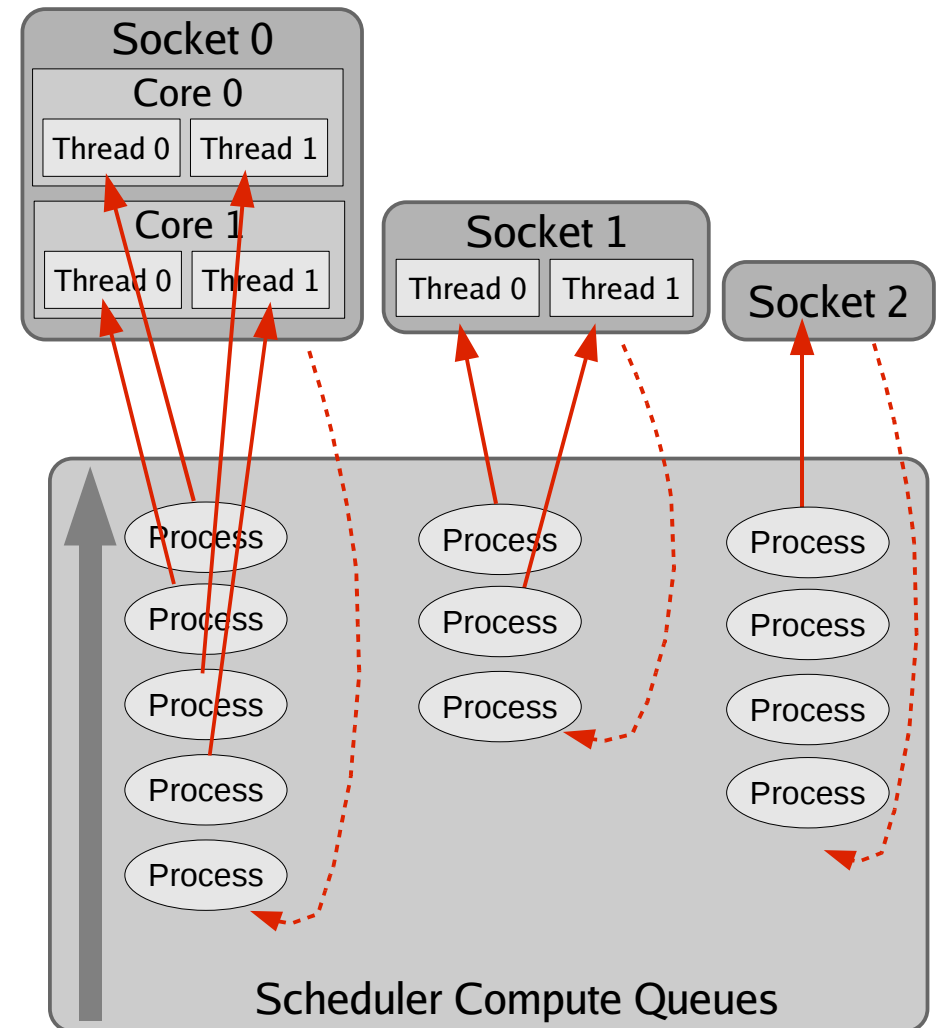
```
Zone:DMA freepages: 2207 min: 0 low: 0 high: 0
Zone:Normal freepages: 484 min: 1279 low: 4544 high: 6304
Zone:HighMem freepages: 266 min: 255 low: 61952 high: 92928
Free pages: 2957 ( 266 HighMem)
( Active: 245828/1297300, inactive_laundry: 194673, inactive_clean: 194668, free: 2957 )
aa:0 ac:0 id:0 il:0 ic:0 fr:2207
aa:630 ac:1009 id:189 il:233 ic:0 fr:484
aa:195237 ac:48952 id:1297057 il:194493 ic:194668 fr:266
1*4kB 1*8kB 1*16kB 1*32kB 1*64kB 0*128kB 0*256kB 1*512kB 0*1024kB 0*2048kB 2*4096kB = 8828kB)
48*4kB 8*8kB 97*16kB 4*32kB 0*64kB 0*128kB 0*256kB 0*512kB 0*1024kB 0*2048kB 0*4096kB =
1936kB)
12*4kB 1*8kB 1*16kB 1*32kB 1*64kB 1*128kB 1*256kB 1*512kB 0*1024kB 0*2048kB 0*4096kB =
1064kB)
Swap cache: add 3838024, delete 3808901, find 107105/1540587, race 0+2
138138 pages of slabcache
1100 pages of kernel stacks
0 lowmem pagetables, 37046 highmem pagetables
Free swap: 3986092kB
4194304 pages of RAM
3833824 pages of HIGHMEM
```

# Tuning File Systems and Disk IO

- Kernel Optimizations
  - CPU Scheduling -multi-threaded, multi-core
  - NUMA – optimized w/ NUMActl
  - Kernel disk I/O – I/O schedulers, Direct I/O, Async I/O
  - File systems EXT3, NFS, GFS, OCFS
  - Database characteristics
  - Huge Pages – Hugetlbfs, db's java etc
- RHEL5 Performance Features

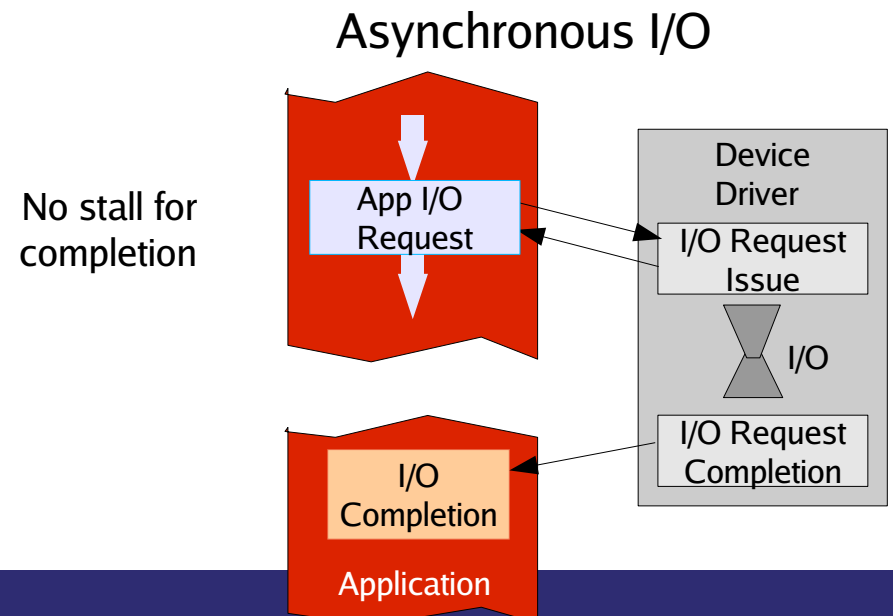
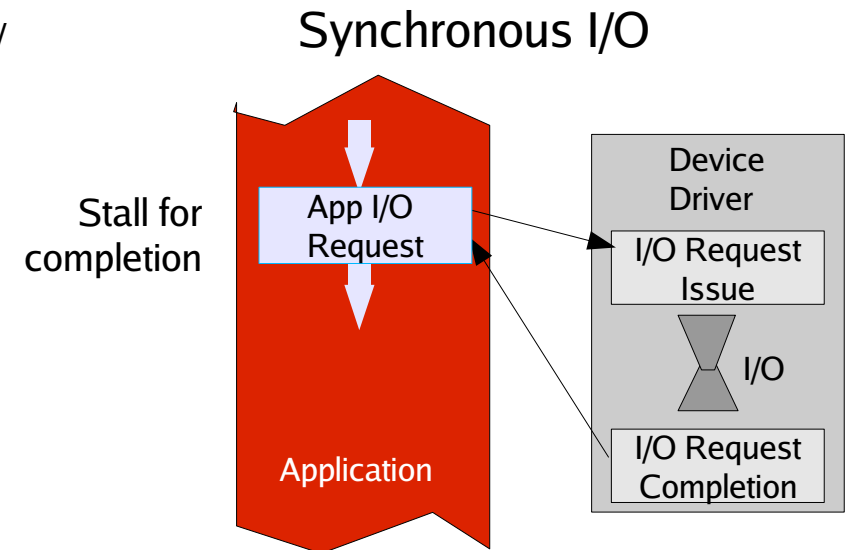
# Linux at 16cpus - quad-core and beyond

- Recognizes differences between logical and physical processors
  - I.E. Multi-core, hyperthreaded & chips/sockets
- Optimizes process scheduling to take advantage of shared on-chip cache, and NUMA memory nodes
- Implements multilevel run queues for sockets and cores (as opposed to one run queue per processor or per system)
  - Strong CPU affinity avoids task bouncing
  - Requires system BIOS to report CPU topology correctly



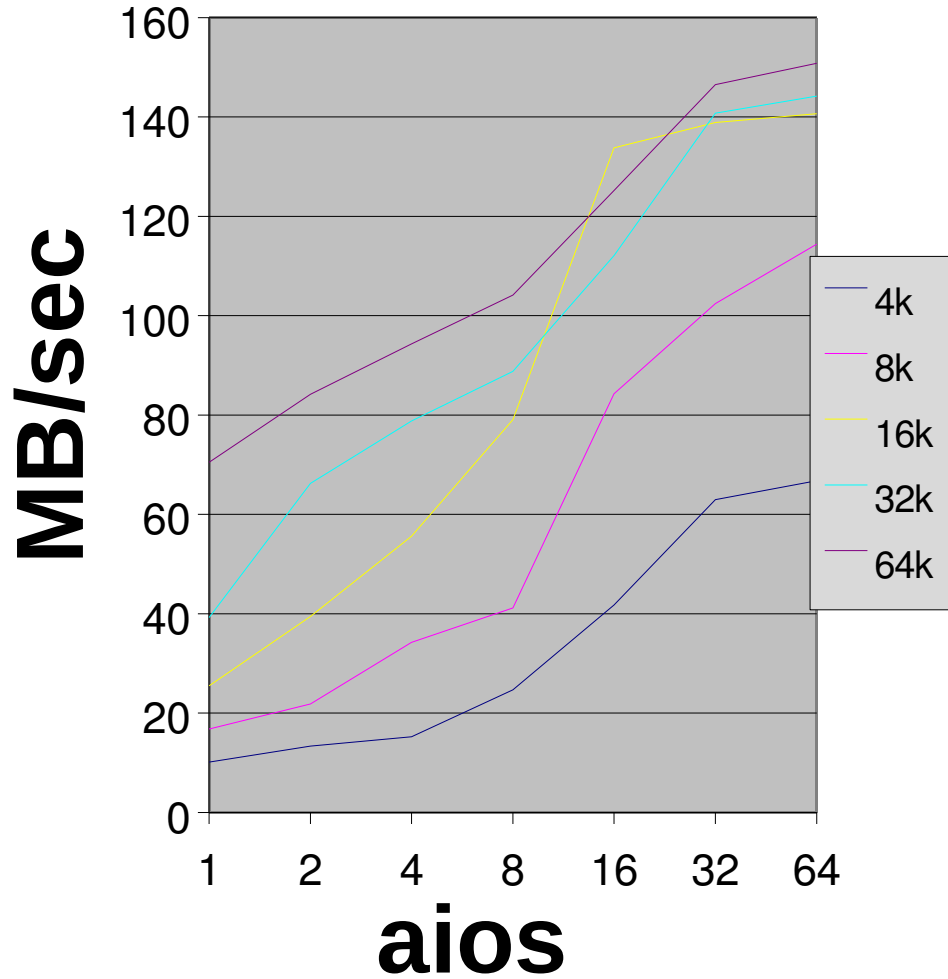
# Asynchronous I/O to File Systems

- Allows application to continue processing while I/O is in progress
  - Eliminates Synchronous I/O stall
  - Critical for I/O intensive server applications
- Red Hat Enterprise Linux – since 2002
  - Support for RAW devices only
- With Red Hat Enterprise Linux 4, significant improvement:
  - Support for Ext3, NFS, GFS file system access
  - Supports Direct I/O (e.g. Database applications)
  - Makes benchmark results more appropriate for real-world comparisons

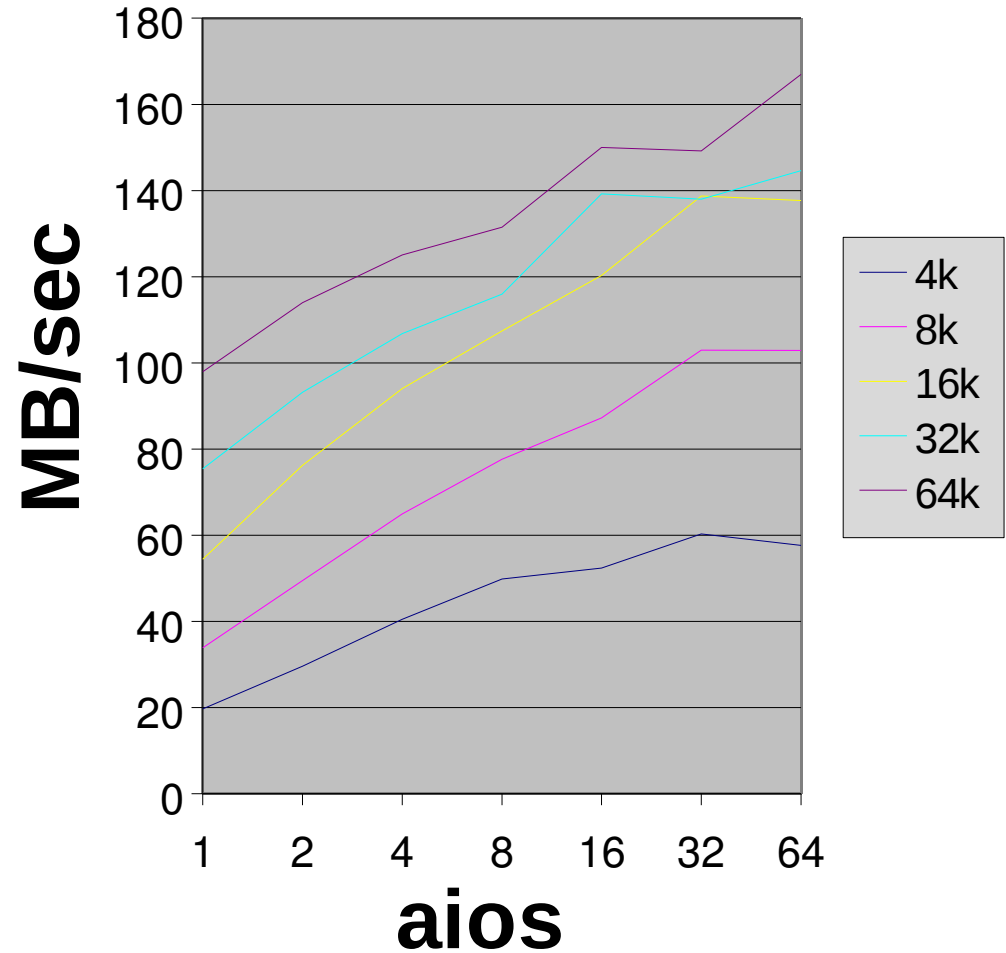


# Asynchronous I/O Characteristics

## R4 U4 FC AIO Read



## R4 U4 FC AIO Write Perf



# Performance Tuning – DISK RHEL3

```
[root@dhcp83-36 sysctl]# /sbin/elvtune /dev/hda
```

```
/dev/hda elevator ID          0
    read_latency:             2048
    write_latency:            8192
    max_bomb_segments:        6
```

```
[root@dhcp83-36 sysctl]# /sbin/elvtune -r 1024 -w 2048 /
dev/hda
```

```
/dev/hda elevator ID          0
    read_latency:             1024
    write_latency:            2048
    max_bomb_segments:        6
```

# Disk IO tuning - RHEL4/5

- RHEL4/5 – 4 tunable I/O Schedulers
  - CFQ – elevator=cfq. Completely Fair Queuing default, balanced, fair for multiple luns, adaptors, smp servers
  - NOOP – elevator=noop. No-operation in kernel, simple, low cpu overhead, leave opt to ramdisk, raid cntrl etc.
  - Deadline – elevator=deadline. Optimize for run-time-like behavior, low latency per IO, balance issues with large IO luns/controllers (NOTE: current best for FC5)
  - Anticipatory – elevator=as. Inserts delays to help stack aggregate IO, best on system w/ limited physical IO – SATA
- RHEL4 - Set at boot time on command line
- RHEL5 – Change on the fly

# File Systems

- Separate swap and busy partitions etc.
- EXT2/EXT3 – separate talk
  - [http://www.redhat.com/support/wpapers/redhat/ext3/\\*.html](http://www.redhat.com/support/wpapers/redhat/ext3/*.html)
  - Tune2fs or mount options
    - data=ordered – only metadata journaled
    - data=journal – both metadata and data journaled
    - data=writeback – use with care !
    - Setup default block size at mkfs -b XX
- RHEL4/5 EXT3 improves performance
  - Scalability upto 5 M file/system
  - Sequential write by using block reservations
  - Increase file system upto 8TB
- GFS – global file system – cluster file system

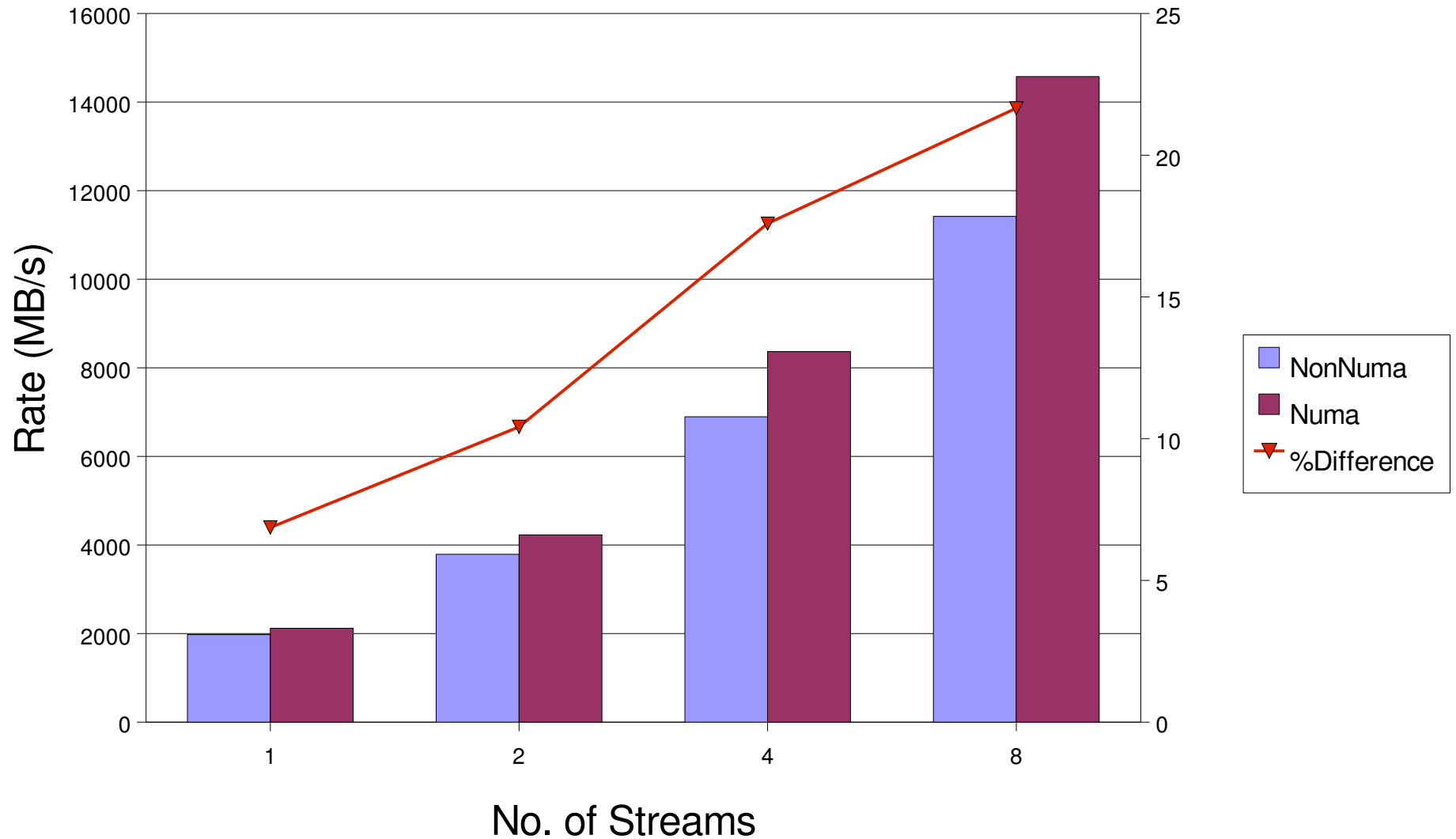
# Optimizing File System Performance

- Use OLTP and DSS workloads
- Results with various database tuning options
  - RAW vs EXT3/GFS/NFS w/ o\_direct (ie directIO in iozone)
  - ASYNC IO options
    - RHEL3 – DIO+AIO not optimal (page cache still active)
    - RHEL4
      - EXT3 supports AIO+DIO out of the box
      - GFS – U2 full support AIO+DIO / Oracle cert
      - NFS – U3 full support of both DIO+AIO
  - HUGHMEM kernels on x86 kernels  
HugeTLBS – use larger page sizes (ipcs)

# Section 4 - Examples

- **General guidelines**
  - **Effect of NUMA and NUMCTL**
  - **Effect CPUspeed howto control**
- **Benchmarking**
  - **McCalpin – know max memory BW**
  - **IOzone – run your own**
- **Database Tuning**
- **JVM Tuning**

# McCalpin Streams Copy Bandwidth (1,2,4,8)



# RHEL4&5 NUMAstat and NUMActl

- NUMAstat to display system NUMA characteristics on a numasystem

```
[root@perf5 ~]# numastat
```

	node3	node2	node1	node0
numa_hit	72684	82215	157244	325444
numa_miss	0	0	0	0
numa_foreign	0	0	0	0
interleave_hit	2668	2431	2763	2699
local_node	67306	77456	152115	324733
other_node	5378	4759	5129	711

- NUMActl to control process and memory”

```
numactl [ --interleave nodes ] [ --preferred node ] [ --membind nodes ]  
[ --cpubind nodes ] [ --localalloc ] command {arguments ...}
```

- **TIP**

- App < memory single NUMA zone
  - Numactl use `–cpubind cpus` within same socket
- App > memory of a single NUMA zone
  - Numactl `–interleave XY` and `–cpubind XY`

# RHEL4&5 NUMAstat and NUMActl

## EXAMPLES

`numactl --interleave=all bigdatabase` arguments Run big database with its memory interleaved on all CPUs.

`numactl --cpubind=0--membind=0,1 process` Run process on node 0 with memory allocated on node 0 and 1.

`numactl --preferred=1` `numactl --show` Set preferred node 1 and show the resulting state.

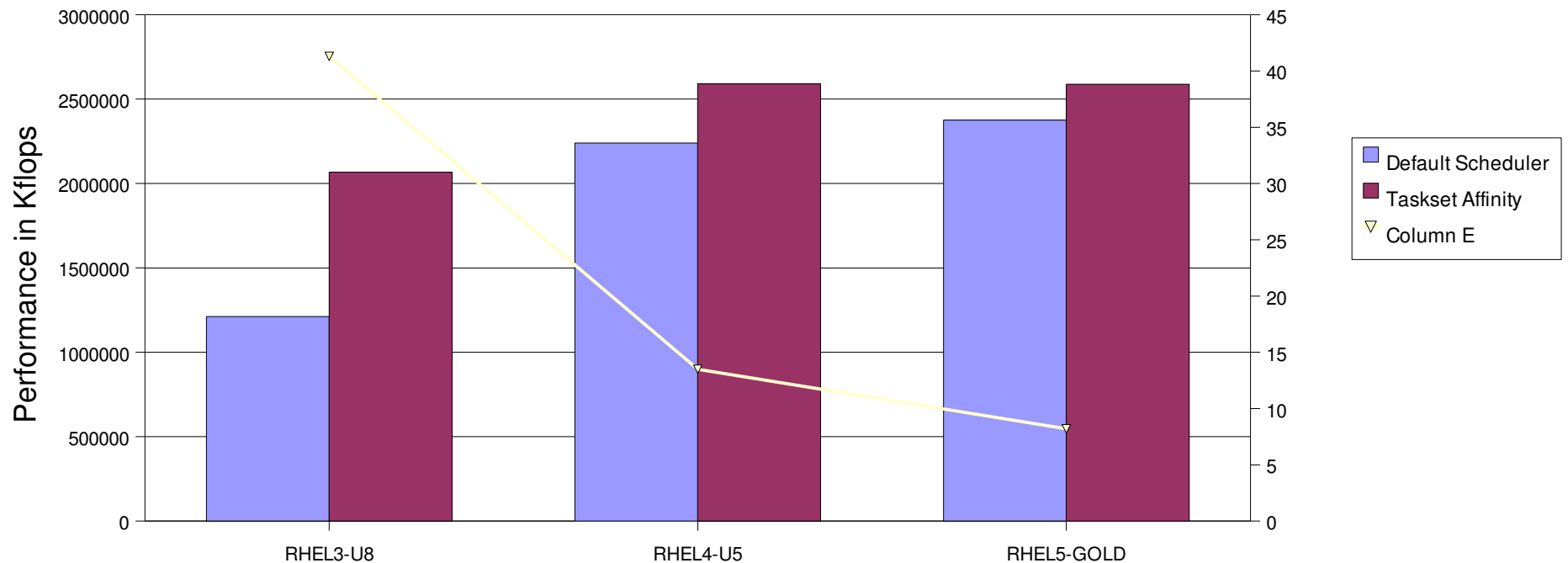
`numactl --interleave=all --shmkeyfile /tmp/shmkey` Interleave all of the sysv shared memory region specified by /tmp/shmkey over all nodes.

`numactl --offset=1G --length=1G --membind=1 --file /dev/shm/A --touch` Bind the second gigabyte in the tmpfs file /dev/shm/A to node 1.

`numactl --localalloc /dev/shm/file` Reset the policy for the shared memory file file to the default localalloc policy.

# Linux NUMA Evolution

RHEL3, 4 and 5 Linpack Multi-stream  
AMD64, 8cpu - dualcore (1/2 cpus loaded)



## ■ Limitations :

- Numa “spill” to different numa boundaries
- Process migrations – no way back
- Lack of page replication – text, read mostly

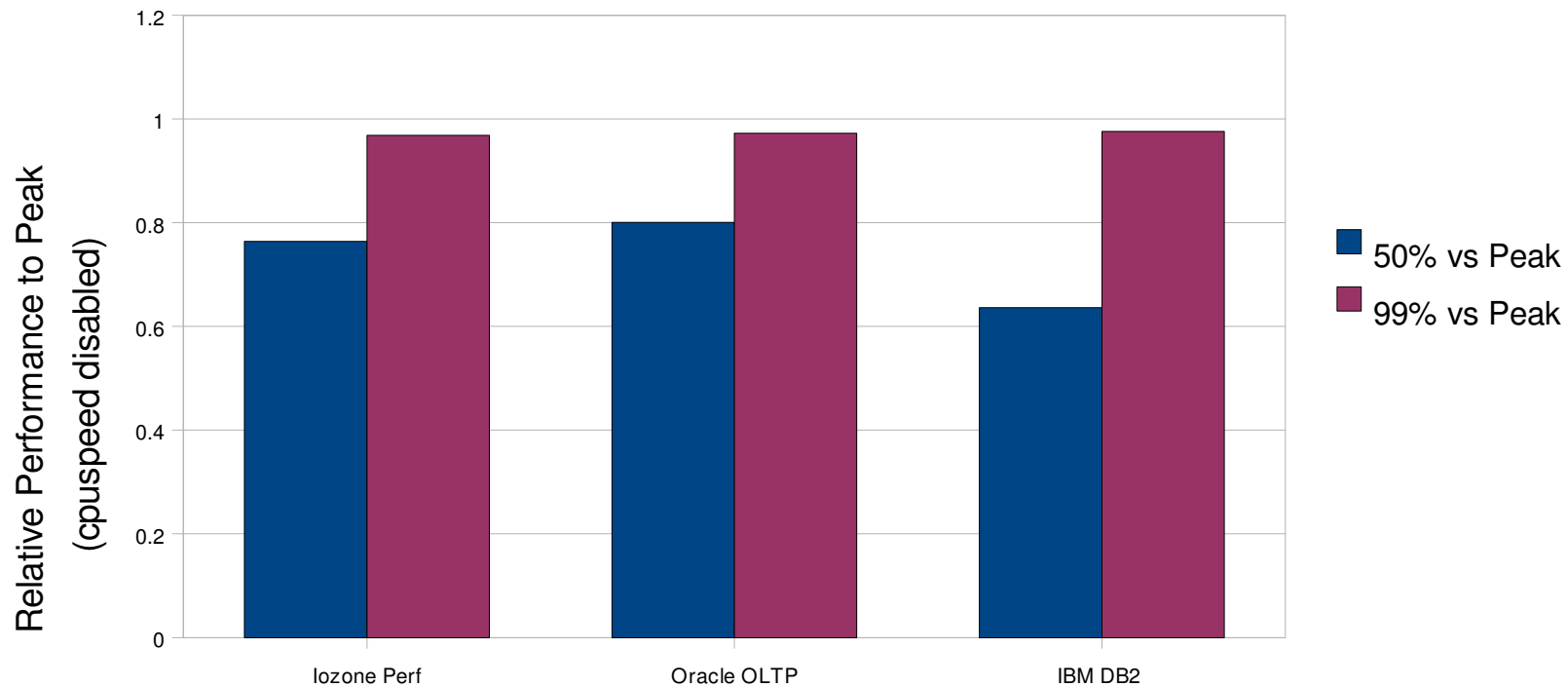
# RHEL5.2 CPU speed and performance:

- Enabled = governor set to “ondemand”
- Looks at cpu usage to regulate power
  - Within 3-5% of performance for cpu loads
  - IO loads can keep cpu stepped down -15-30%
- Supported in RHEL5.2 virtualization
- To turn off – else may leave cpu’s in reduced step
  - If its not using performance, then:
  - `# echo performance > /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor`
  - Then check to see if it stuck:
  - `# cat /sys/devices/system/cpu/cpu0/cpufreq/scaling_governor`
  - Check /proc/cpuinfo to make sure your seeing the expected CPU freq.
- Proceed to “normal” service disable
  - Service cpuspeed stop
  - Chkconfig cpuspeed off

# Effects of CPU speed to peak performance:

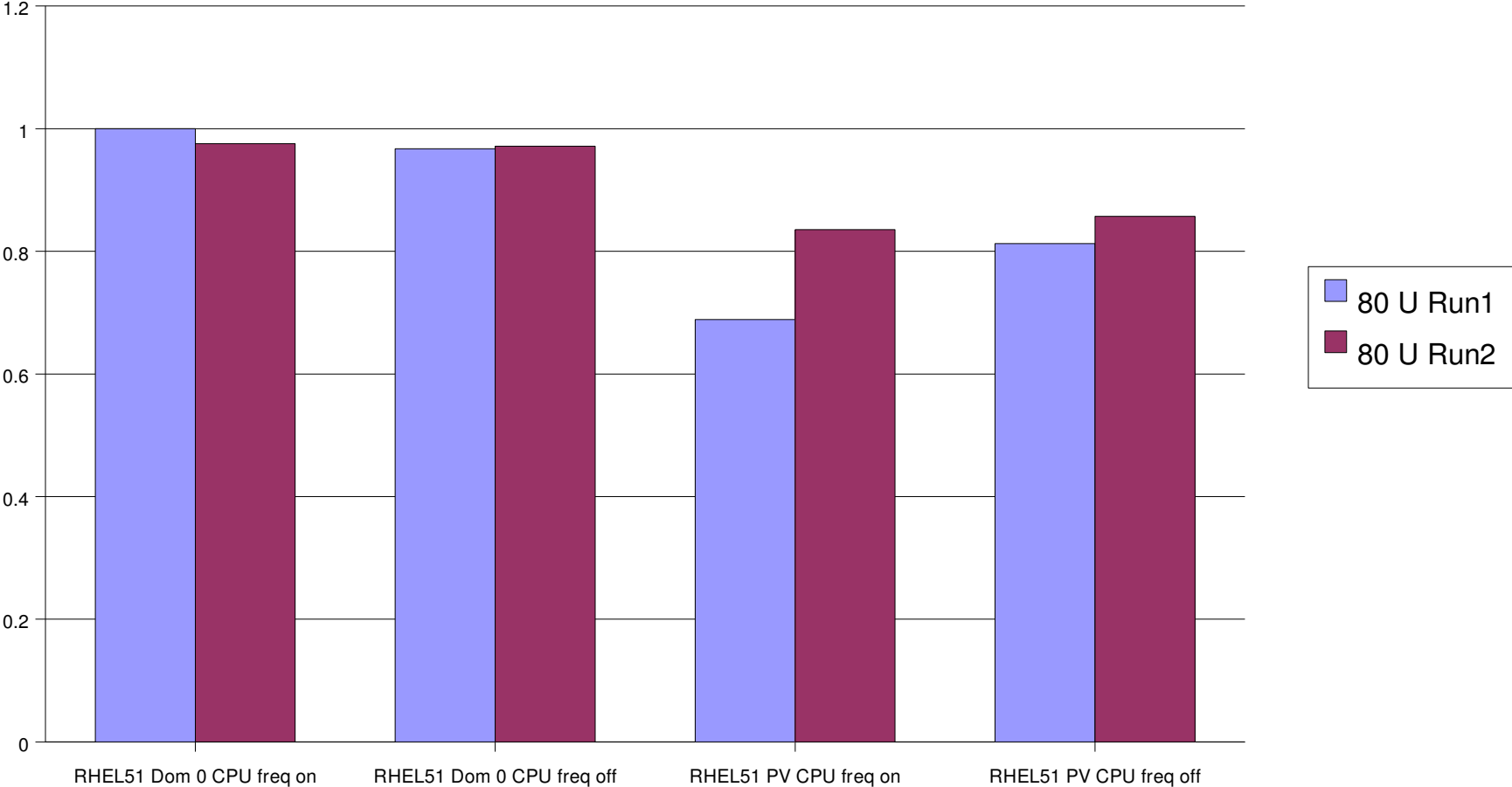
## RHEL5.2 Effect of CPU speed on I/O workloads

Intel 4-cpu, 16 Gb memory, FC disk



# Effects of CPU speed with RHEL5.2 Virtualization

Oracle runs with CPU Freq Xen kernel

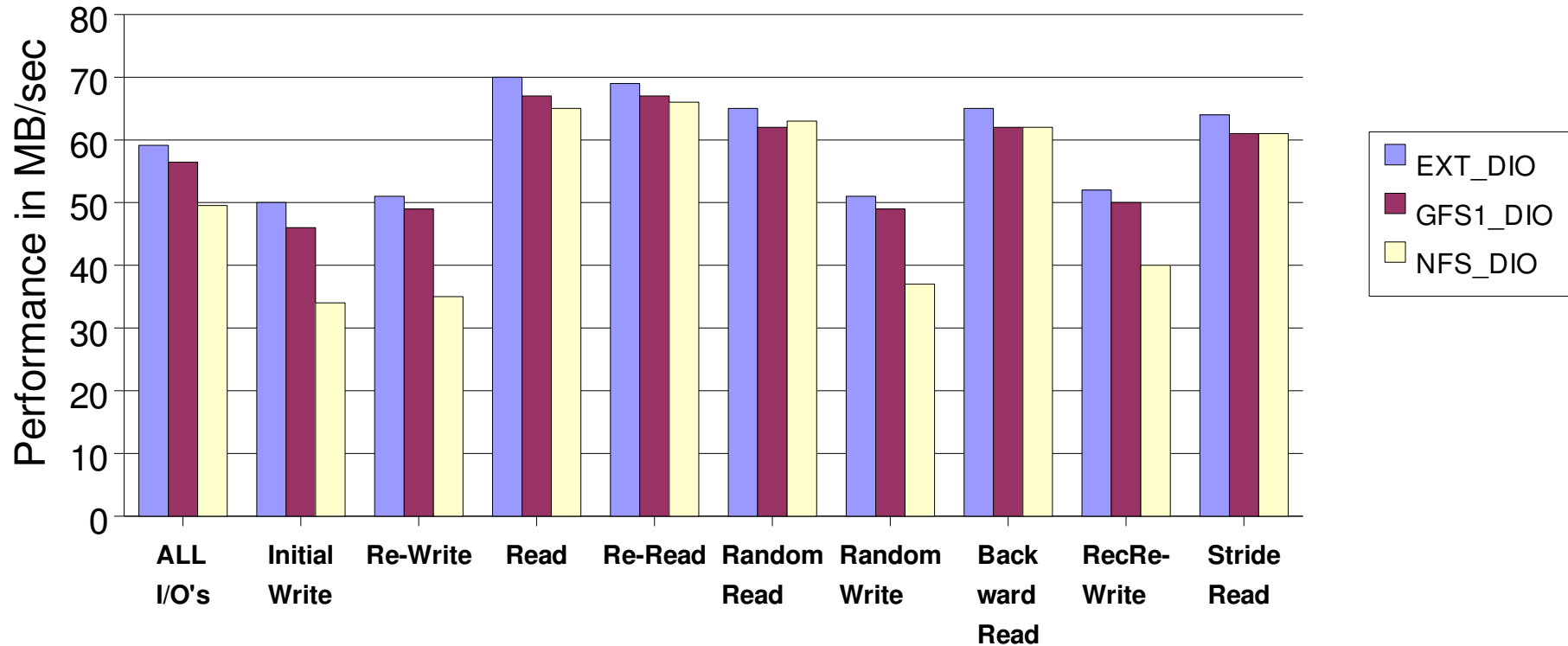


# Using IOzone w/ o\_direct – mimic database

- **Problem :**
  - **Filesystems use memory for file cache**
  - **Databases use memory for database cache**
  - **Users want filesystem for management outside database access (copy, backup etc)**
- **You DON'T want BOTH to cache.**
- **Solution :**
  - **Filesystems that support Direct IO**
  - **Open files with o\_direct option**
  - **Databases which support Direct IO (ORACLE)**
  - **NO DOUBLE CACHING!**

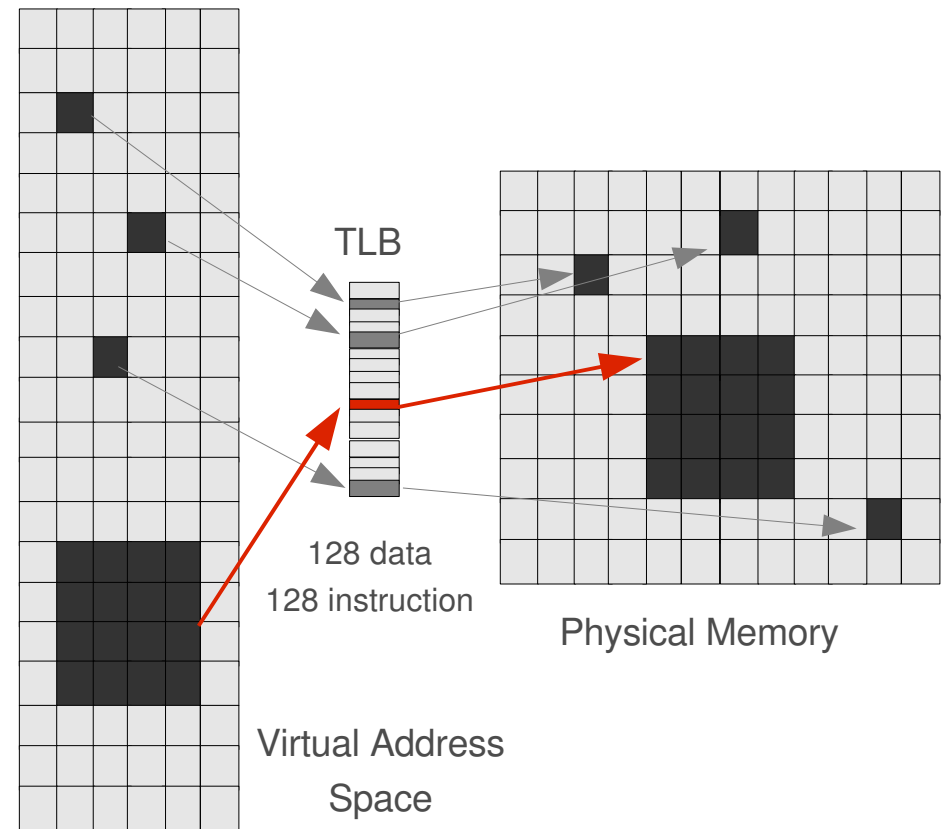
# EXT3, GFS, NFS I/Ozone w/ DirectIO

RHEL5 Direct\_IO IOzone EXT3, GFS, NFS  
(Geom 1M-4GB, 1k-1m)



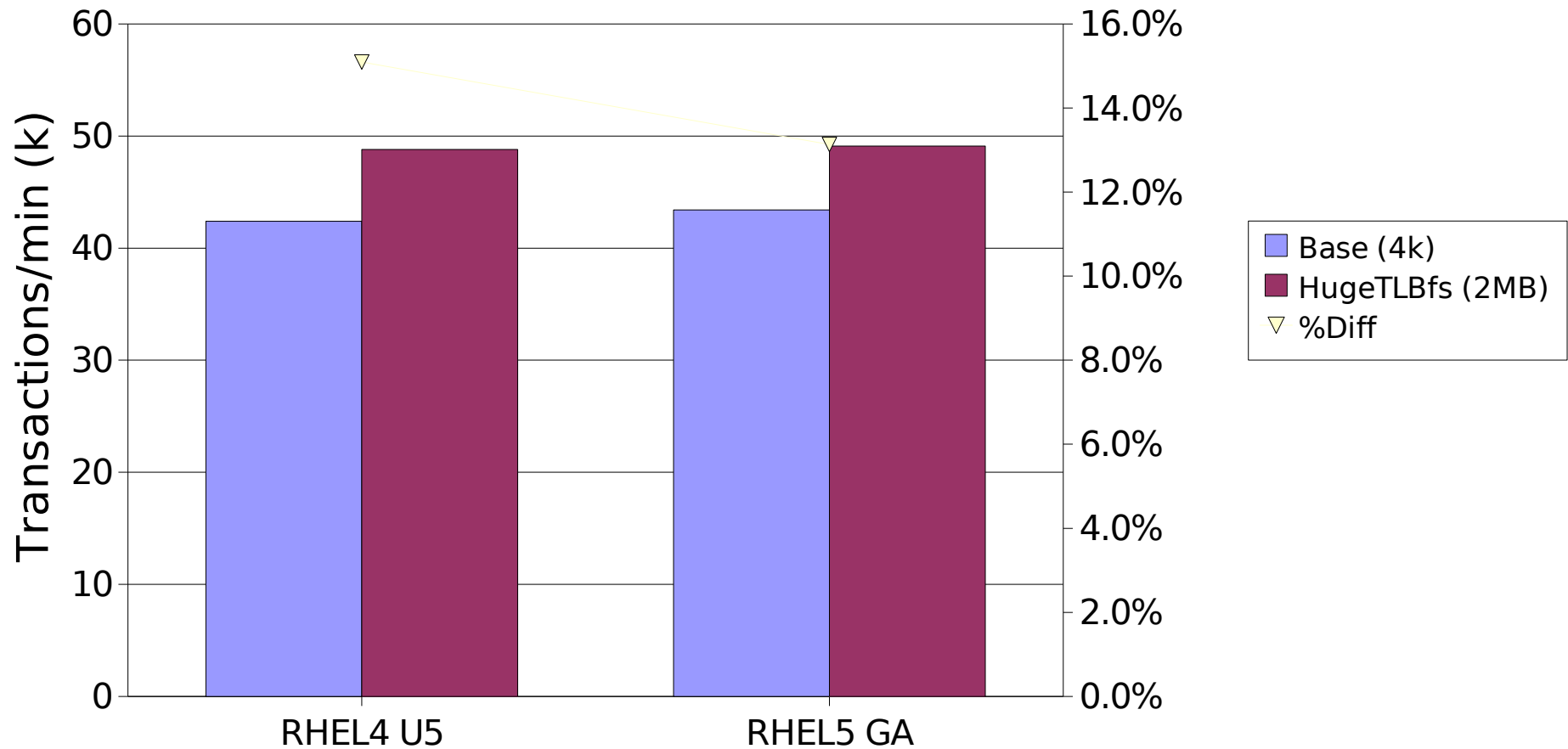
# HugeTLBFS

- The Translation Lookaside Buffer (TLB) is a small CPU cache of recently used virtual to physical address mappings
- TLB misses are extremely expensive on today's very fast, pipelined CPUs
- Large memory applications can incur high TLB miss rates
- HugeTLBs permit memory to be managed in very large segments
  - E.G. Itanium®:
    - Standard page: 16KB
    - Default huge page: 256MB
    - 16000:1 difference
- File system mapping interface
- Ideal for databases
  - E.G. TLB can fully map a 32GB Oracle SGA



# Using HugeTLBfs w/ Databases

RHEL4+5 Effect of HugeTLBfs  
Oracle 10G OLTP Performance  
Intel 4cpu, 8GB memory, FC San



# JVM Tuning

- Eliminate swapping
  - Lower swappiness to 10%(or lower if necessary).
- Promote pagecache reclaiming
  - Lower dirty\_background\_ratio to 10%
  - Lower dirty\_ratio if necessary
- Promote inode cache reclaiming
  - Lower vfs\_cache\_pressure

# Tuning Network Apps Messages/sec

- **Disable cpuspeed, selinux, auditd, irqbalance**
- **Manual binding IRQs w/ multiple nics**
  - echo values > /proc/irq/XXX or use “TUNA”
  - Intel ixgb IRQs send/recv to cpu socket w/ shared cache
- **Use Taskset -c to start applications on**
  - 1 cpu per socket – good for BW intensive app
- **Shield cpus for critical apps**
  - Move all existing processes off of the core(s) to cpu0
- **Pairs of cpus on the same socket – shared 2nd level cache**
- **Keep user apps on cpus separate from Network apps**

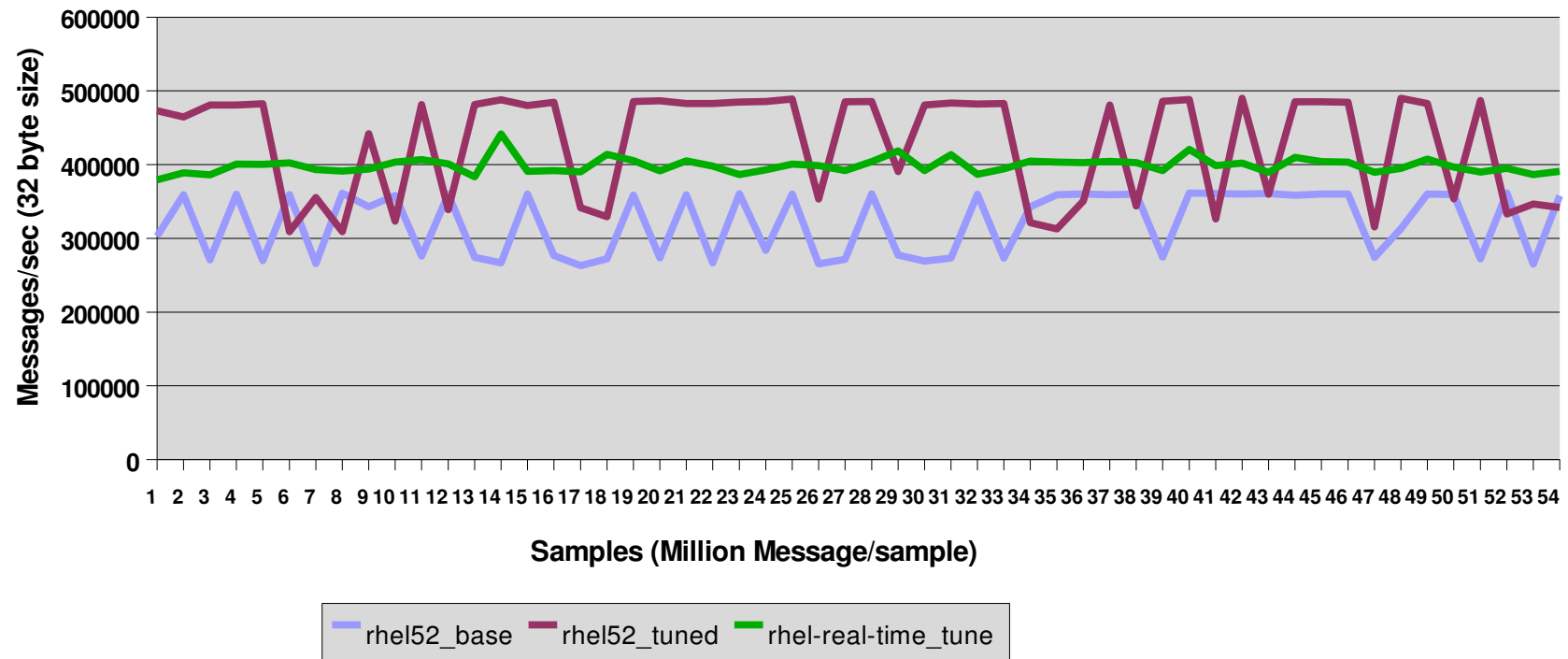
# RT Tuning Network Apps Messages/sec

10 Gbit Nics Stoakley 2.67 to Bensley 3.0 Ghz

Tuning enet gains +25% in Ave Latency,

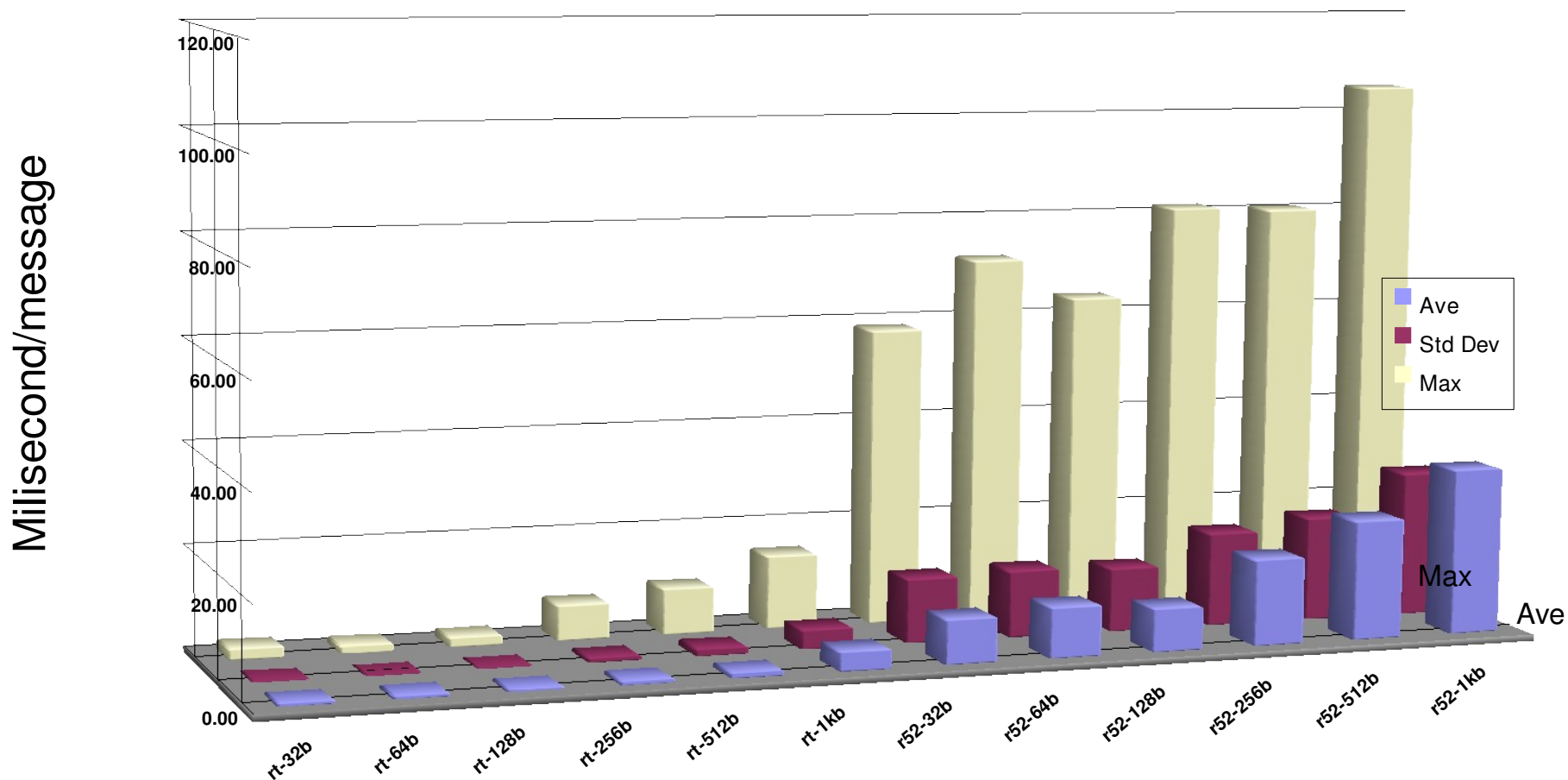
RT kernel reduced peak latency but smoother – how much?

Red Hat MRG Performance AMQP Mess/s  
Intel 8-cpu/16gb, 10Gb enet



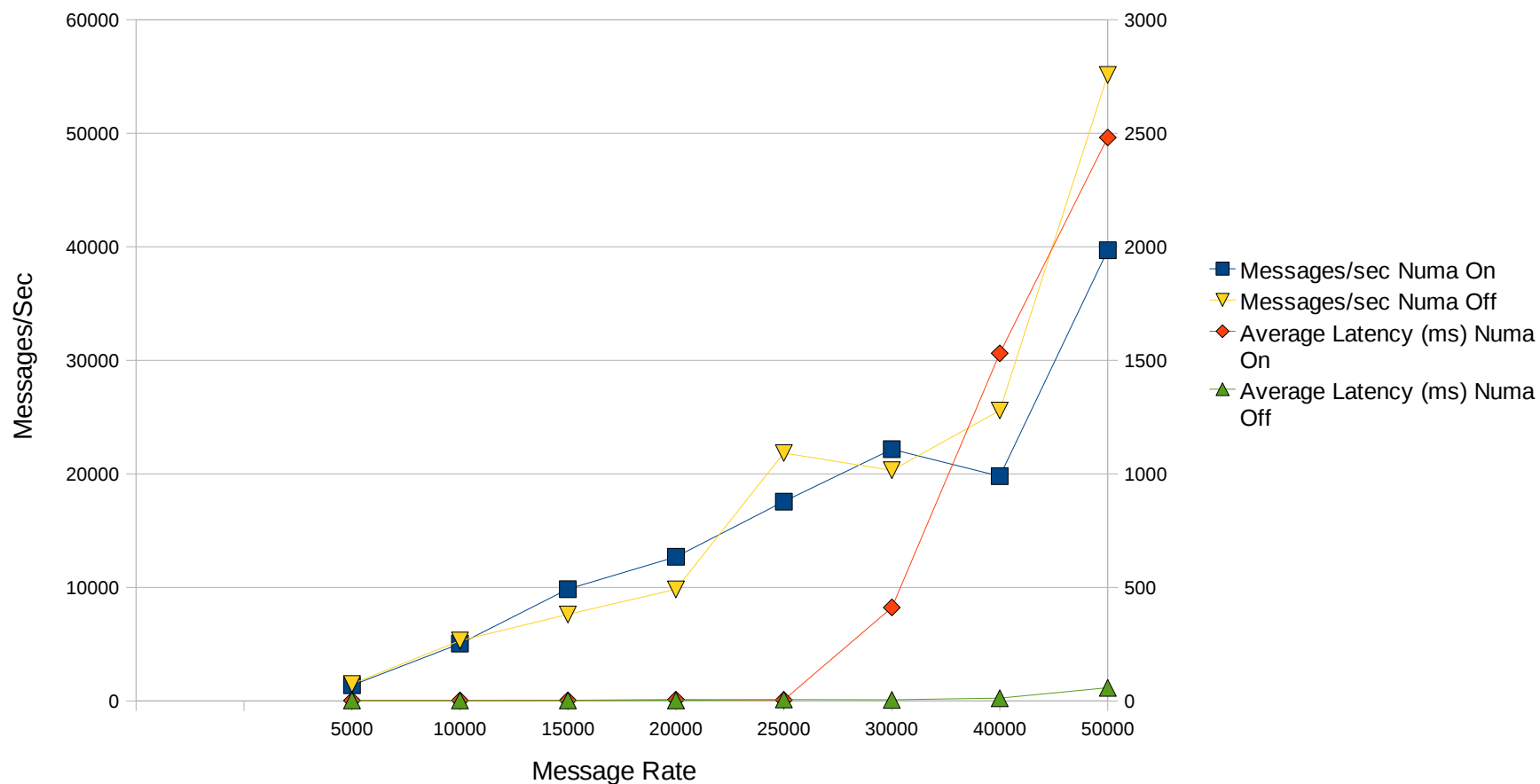
# RT Performance of Network Apps Messages/sec

RH AMQP Latency on Intel 8cpu/10Gbit enet  
RHEL5.2 and RHEL-RT



# Numa Network Apps Messages/sec

Wombat Messages/sec RHEL5.2  
Effects with Numa On/Off



# General Performance Tuning Guidelines

- **Use hugepages whenever possible.**
- **Minimize swapping.**
- **Maximize pagecache reclaiming**
- **Place swap partition(s) on quite device(s).**
- **Direct IO if possible.**
- **Beware of turning NUMA off.**

# Benchmark Tuning

- **Use Hugepages.**
- **Dont overcommit memory**
- **If memory must be over committed**
  - **Eliminate all swapping.**
  - **Maximize pagecache reclaiming**
  - **Place swap partition(s) on separate device(s).**
- **Use Direct IO**
- **Dont turn NUMA off.**

# Linux Performance Tuning References

- Alikins, ?System Tuning Info for Linux Servers,  
[http://people.redhat.com/alikins/system\\_tuning.html](http://people.redhat.com/alikins/system_tuning.html)
- Axboe, J., ?Deadline IO Scheduler Tunables, SuSE, EDF R&D, 2003.
- Braswell, B, Ciliendo, E, ?Tuning Red Hat Enterprise Linux on IBM eServer xSeries Servers, <http://www.ibm.com/redbooks>
- Corbet, J., ?The Continuing Development of IO Scheduling?,  
<http://lwn.net/Articles/21274>.
- Ezolt, P, Optimizing Linux Performance, [www.hp.com/hpbooks](http://www.hp.com/hpbooks), Mar 2005.
- Heger, D, Pratt, S, ?Workload Dependent Performance Evaluation of the Linux 2.6 IO Schedulers?, Linux Symposium, Ottawa, Canada, July 2004.
- Red Hat Enterprise Linux “Performance Tuning Guide”  
[http://people.redhat.com/dshaks/rhel3\\_perf\\_tuning.pdf](http://people.redhat.com/dshaks/rhel3_perf_tuning.pdf)
- Network, NFS Performance covered in separate talks  
<http://nfs.sourceforge.net/nfs-howto/performance.html>

# Questions?

