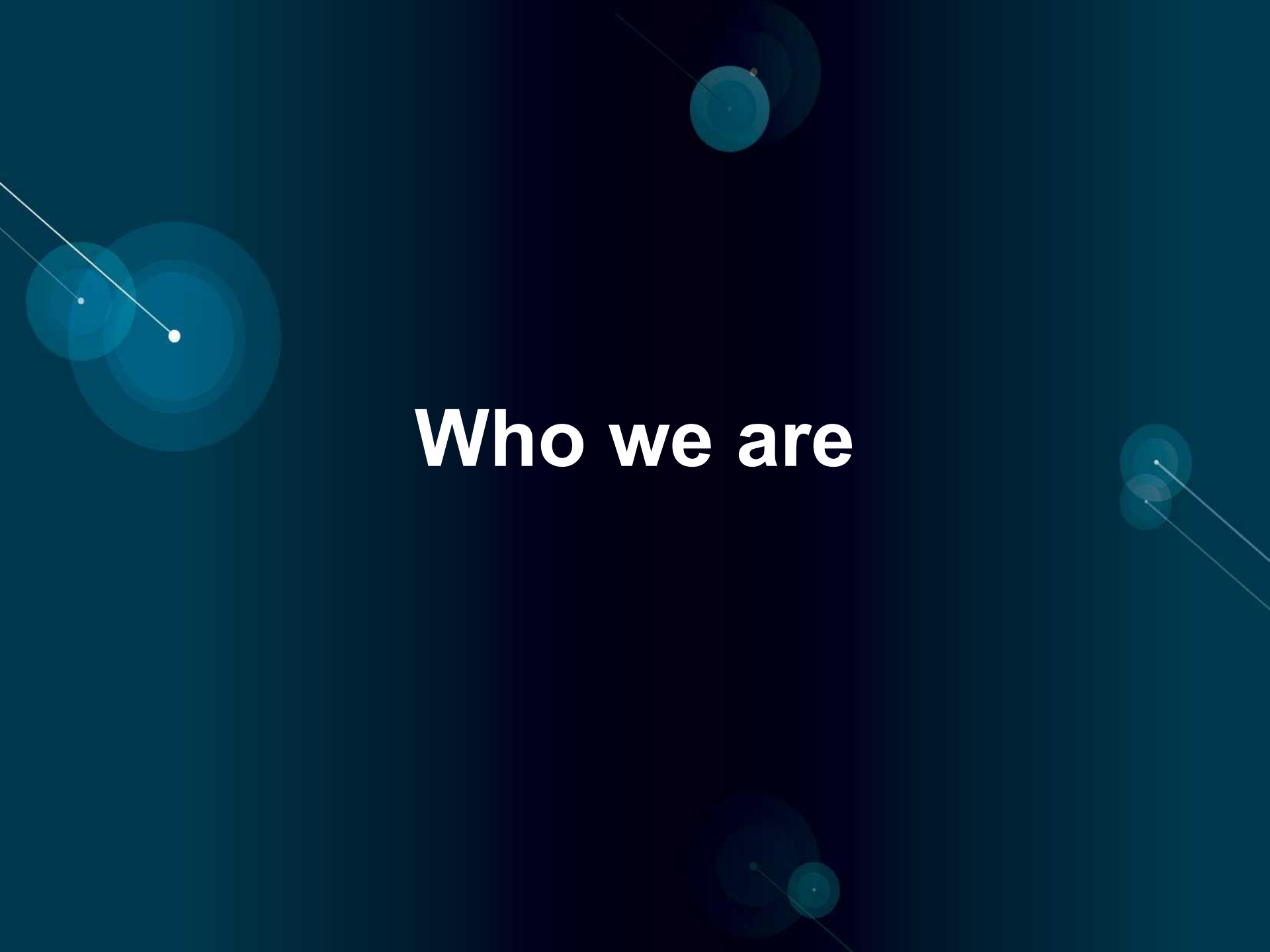


Performance Analysis Techniques from Red Hat

@jeremyeder

Agenda

- Who we are
- Approach to Performance Analysis
- Evolution of Performance in RHEL
- Subsystem Analysis
- Tuned
- Future Plans



Who we are

Red Hat Performance Engineering

Characterize Performance

- RHEL + Addons
- RHEV
- JBoss
- Gluster & Ceph
- MRG
- CloudForms
- OpenShift/OpenStack
- Latest Hardware
- Many more...

Enhance Subscription

- Enable Partners
- Pioneer new tech
- Support critical customer issues
- Participate in Benchmark Consortia
- Summit Presos
- TAM Webinars

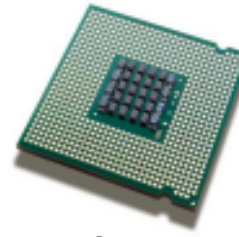
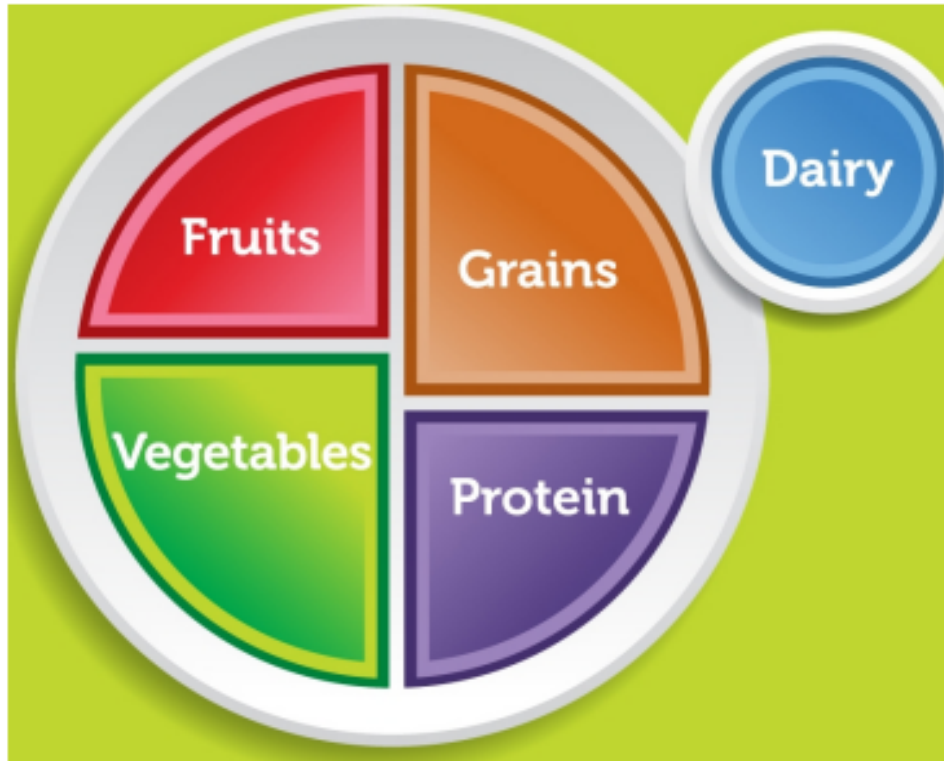
Partner with Engineering teams

- Evaluate experimental patches, upstream backports [slab/slub, pre-warm migrate, loopback bypass]
- Provide feedback to developers on what's working and what's not
- Provide guidance in early planning/dev phase, embedded in team [kvm, RHS, OpenShift]



Approach to Performance Analysis

Approach to Performance Analysis



linpack, streams, sysbench, libmicro



iozone, smallfile
sysbench



netperf

Approach to Performance Analysis

- Test for limits
- Quantify overheads
- See what breaks
- Assemble/ship and document tuning advice
- Adapt tests to container-friendly patterns

Evolution of Performance in RHEL

RHEL5

Static Hugepages

Ktune – on/off

CPU Affinity (taskset)

NUMA Pinning (numactl)

irqbalance



RHEL6

Transparent Hugepages

Tuned – choose profile

CPU Affinity (ts/numactl)

NUMAD – userspace tool

irqbalance – NUMA enhanced



RHEL7

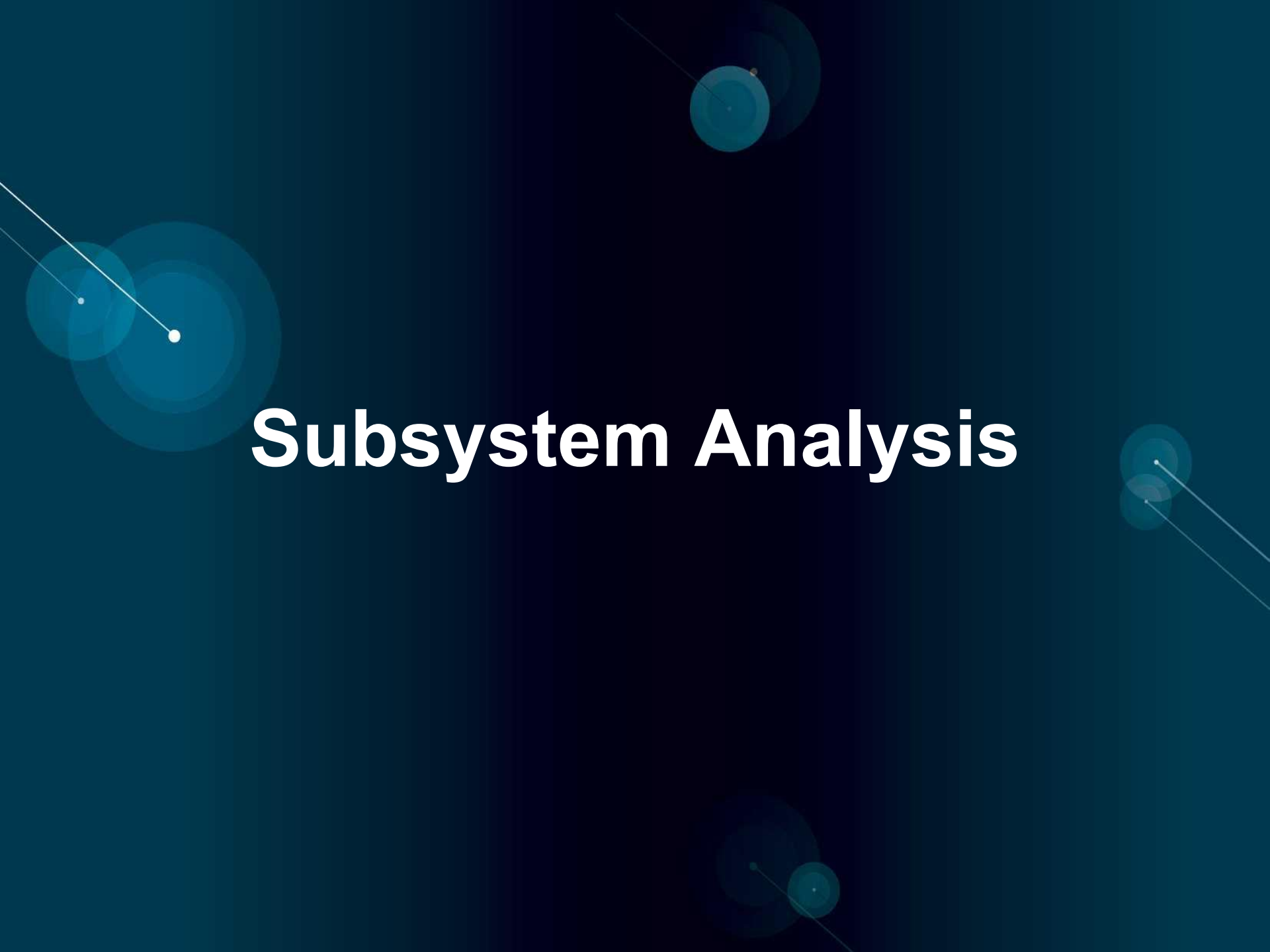
Transparent Hugepages

Tuned – throughput-performance (default)

CPU Affinity (ts/numactl)

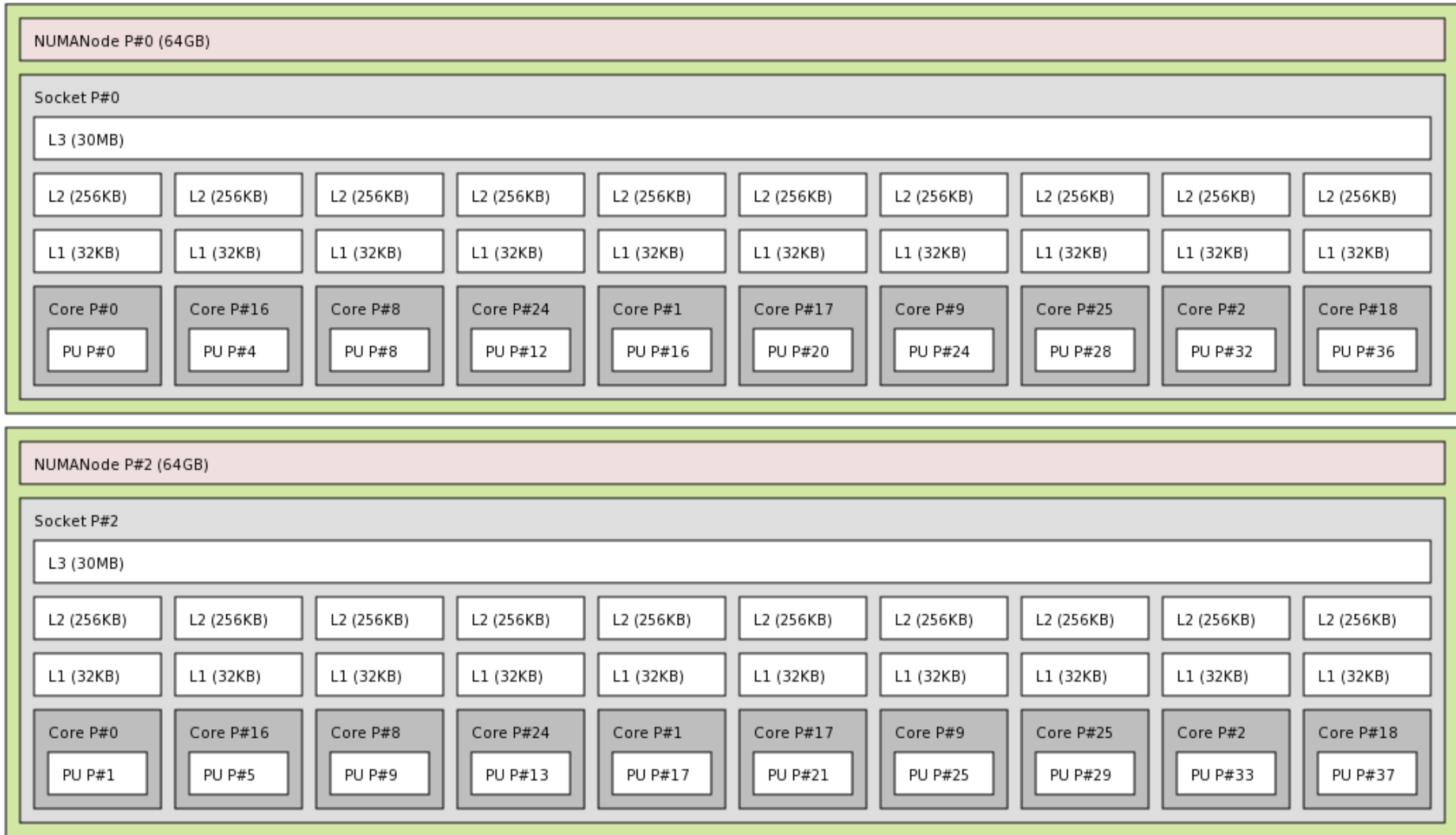
Autonuma-Balance

irqbalance – NUMA enhanced

The background is a dark teal gradient. It features several abstract elements: a large cluster of overlapping teal circles on the left, a smaller cluster of two circles at the top center, a cluster of three circles on the right, and a cluster of two circles at the bottom center. Thin white lines extend from the circles, some ending in small white dots.

Subsystem Analysis

NUMA Topology: Istopo



Subsystem Analysis: Disk

```
# iostat -N 1
Linux 3.10.0-229.el7.x86_64 (jermis-lab7.perf.lab.eng.rdu.redhat.com)
    03/12/2015    _x86_64_ (24 CPU)
```

avg-cpu:	%user	%nice	%system	%iowait	%steal	%idle
	0.03	0.00	0.01	0.00	0.00	99.95

Device:	tps	kB_read/s	kB_wrtn/s	kB_read	kB_wrtn
vda	5.42	17.19	59.87	19633224	68398872
vdb	0.46	0.98	5.48	1119751	6263272
vg0-swap	0.00	0.00	0.00	1008	0
vg0-root	5.46	17.18	59.87	19625293	68396764

avg-cpu:	%user	%nice	%system	%iowait	%steal	%idle
	0.04	0.00	0.71	3.31	0.00	95.94

Device:	tps	kB_read/s	kB_wrtn/s	kB_read	kB_wrtn
vda	7043.00	0.00	12326.00	0	12326
vdb	0.00	0.00	0.00	0	0
vg0-swap	0.00	0.00	0.00	0	0
vg0-root	7042.00	0.00	12322.00	0	12322

Subsystem Analysis: Memory

```
# vmstat 1
```

procs		-----memory-----				---swap--		-----io-----		-system--		-----cpu-----				
r	b	swpd	free	buff	cache	si	so	bi	bo	in	cs	us	sy	id	wa	st
2	0	0	25554152	986896	13833932	0	0	0	0	0	307	266	0	1	99	0
0																
1	0	0	24454408	986896	13833932	0	0	0	0	0	1286	548	0	4	96	0
0																
1	0	0	23428300	986896	13833932	0	0	0	0	0	1288	508	0	4	96	0
0																
1	0	0	22371768	986896	13833932	0	0	0	0	8	1120	150	0	4	96	0
0																
1	0	0	21571872	986896	13833932	0	0	0	0	0	1162	305	1	4	96	0
0																
1	0	0	21571872	986896	13833932	0	0	0	0	0	1040	67	4	0	96	0
0																
1	0	0	21571872	986896	13833932	0	0	0	0	0	1067	90	4	0	96	0
0																
1	0	0	21571872	986896	13833932	0	0	0	0	0	1045	70	4	0	96	0
0																
0	0	0	25773696	986896	13833932	0	0	0	0	24	487	130	2	0	98	0
0																

Subsystem Analysis: Memory

```
# numastat -c qemu Per-node process memory usage (in Mbs)
```

PID		Node 0	Node 1	Node 2	Node 3	Total
10587	(qemu-kvm)	1216	4022	4028	1456	10722
10629	(qemu-kvm)	2108	56	473	8077	10714
10671	(qemu-kvm)	4096	3470	3036	110	10712
10713	(qemu-kvm)	4043	3498	2135	1055	10730
Total		11462	11045	9672	10698	42877

UNALIGNED

```
# numastat -c qemu
```

```
Per-node process memory usage (in Mbs)
```

PID		Node 0	Node 1	Node 2	Node 3	Total
10587	(qemu-kvm)	0	10723	5	0	10728
10629	(qemu-kvm)	0	0	5	10717	10722
10671	(qemu-kvm)	0	0	10726	0	10726
10713	(qemu-kvm)	10733	0	5	0	10738
Total		10733	10723	10740	10717	42913

ALIGNED

Subsystem Analysis: Memory

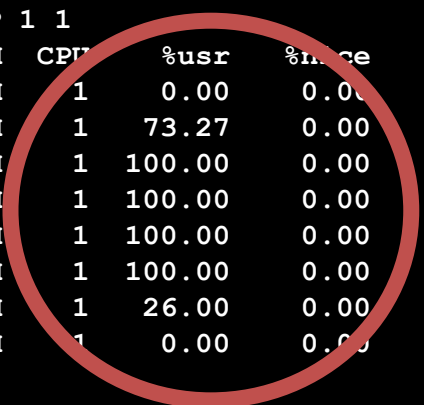
```
# numastat -mczs
```

```
Per-node system memory usage (in MBs):
```

	Node 0	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Total
	-----	-----	-----	-----	-----	-----	-----	-----	-----
MemTotal	32766	32768	32768	32768	32768	32768	32768	32752	262126
MemFree	31863	31965	32120	32086	32098	32080	32114	32062	256388
MemUsed	903	803	648	682	670	688	654	690	5738
FilePages	11	26	8	37	21	18	9	45	176
Slab	25	16	7	10	12	36	10	10	126
Active	5	13	4	25	10	9	6	41	113
Active(file)	4	11	3	23	8	6	3	40	99
SUnreclaim	19	10	6	6	9	33	7	7	97
Inactive	7	15	4	14	12	12	6	6	76
Inactive(file)	7	15	4	14	12	12	6	6	76
SReclaimable	7	6	2	4	3	3	3	2	29
Active(anon)	2	1	1	2	2	2	3	2	14
AnonPages	2	1	1	2	2	2	3	2	14
Mapped	0	0	0	1	4	3	1	1	11
KernelStack	9	0	0	0	0	0	0	0	10
PageTables	0	0	0	0	1	1	0	1	3
Shmem	0	0	0	0	0	0	0	0	0
Inactive(anon)	0	0	0	0	0	0	0	0	0

Subsystem Analysis: CPU

```
# mpstat -P 1 1
04:27:56 PM CPU0  %usr  %nice  %sys %iowait  %irq  %soft  %steal  %guest  %gnice  %idle
04:27:57 PM 1     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00  100.00
04:27:58 PM 1    73.27   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   26.73
04:27:59 PM 1   100.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
04:28:00 PM 1   100.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
04:28:01 PM 1   100.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
04:28:02 PM 1   100.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00
04:28:03 PM 1    26.00   0.00   1.00   0.00   0.00   0.00   0.00   0.00   0.00   73.00
04:28:04 PM 1     0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00   0.00  100.00
```



Subsystem Analysis: Network

```
3.10.0-231.el7.bz1200859.x86_64, p1p1 (sfc 10000Mbit/s link:yes), t=1000ms, cpus=5+1/16  
(consider to increase your sampling interval, e.g. -t 10000)
```

rx:	339.255 MiB/t	443006 pkts/t	668 drops/t	0 errors/t
tx:	10.589 MiB/t	91660 pkts/t	0 drops/t	0 errors/t

rx:	815972.271 MiB	1261075090 pkts	1633625 drops	0 errors
tx:	859607.086 MiB	1308314479 pkts	0 drops	0 errors

sys:	94436 cs/t	1569 procs	2 running	0 iowait
------	------------	------------	-----------	----------

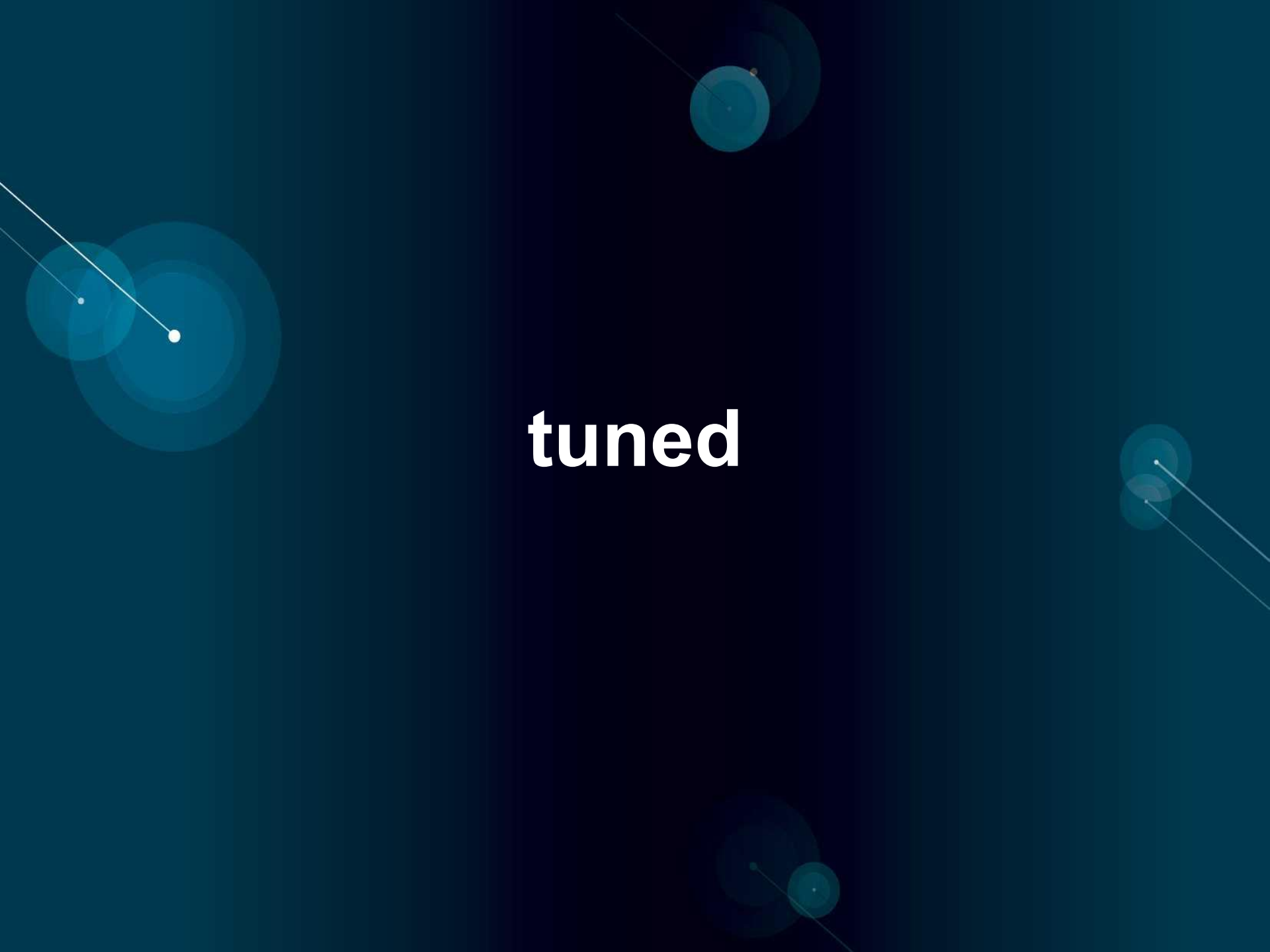
mem:	128728M total	4109M used	2062M active	521M inactive
swap:	1023M total	0M used	0M cached	

cpu13 +:	1.0% usr/t	15.5% sys/t	83.5% idl/t	0.0% iow/t
cpu15 :	1.0% usr/t	13.5% sys/t	85.4% idl/t	0.0% iow/t
cpu 1 :	1.0% usr/t	12.4% sys/t	86.6% idl/t	0.0% iow/t
cpu 3 :	1.0% usr/t	12.1% sys/t	86.9% idl/t	0.0% iow/t
cpu 5 :	2.0% usr/t	10.2% sys/t	87.8% idl/t	0.0% iow/t
cpu14 -:	0.0% usr/t	0.0% sys/t	100.0% idl/t	0.0% iow/t
avg:	0.0%	4.4%	95.6%	0.0%

cpu13 +:	19658 irq/t	19239 irq rx/t	74 irq tx/t
cpu15 :	14126 irq/t	13910 irq rx/t	57 irq tx/t
cpu 9 :	12391 irq/t	12364 irq rx/t	27 irq tx/t
cpu 1 :	11082 irq/t	10996 irq rx/t	33 irq tx/t
cpu 5 :	9962 irq/t	9829 irq rx/t	68 irq tx/t

Subsystem Analysis: ALL

#<-----CPU----->				<-----Disks----->				<-----Network----->				#<-----Memory----->					
#cpu	sys	inter	ctxsw	CPU	KBWrit	Writes	KBIn	PktIn	KBOut	PktOut	#Free	Buff	Cach	Inac	Slab	Map	
0	0	210	179	0	0	64	18	2	17	0	1	32355M	13M	52M	91M	44M	
0	0	202	150	0	0	32			14	0	1	32355M	13M	52M	92M	44M	
4	1	1678	2073	6176	650	108			50	5	33	32346M	14M	57M	98M	44M	
17	0	2348	183	0	0	36	10	2	14	0	3	32346M	14M	57M	98M	44M	
17	0	2361	216	0	0	32	10	1	17			32346M	14M	57M	98M	44M	
7	1	1760	1629	272	20	88356	282	11	63			32345M	14M	58M	96M	44M	
3	2	1691	2526	40	10	795720	2136	0	11			32344M	14M	58M	96M	44M	
3	2	1875	2856	28	7	924736	2714	2	18	0	3	32344M	14M	58M	96M	44M	
2	1	5137	5383	460	40	288836	851	35127	2583	161	2473	32345M	15M	58M	96M	44M	
4	3	16997	28627	0	0	56	10	245172	17629	1101	17088	32344M	15M	58M	96M	44M	
3	2	15619	28062	0	0	44	12	242954	17508	1087	16871	32345M	15M	58M	96M	44M	
6	2	4495	7098	104	3	80	9	51652	3781	210	3675	31804M	15M	59M	96M	44M	
17	5	2380	187	0	0	20	5	1	12	0	3	28287M	15M	59M	96M	44M	
17	5	2349	188	0	0	52	15	1	13	0	1	24805M	15M	59M	96M	44M	
17	5	2356	214	0	0	32	10	2	16	0	1	21284M	15M	59M	96M	44M	
17	5	2348	197	0	0	32	10	0	9	0	1	17436M	15M	59M	96M	44M	
9	3	1366	225	0	0	32	10	2	20	0	4	24766M	15M	59M	96M	44M	
1	0	465	516	8	2	992	169	2	25	1	15	32344M	15M	59M	96M	44M	
1	0	236	205	0	0	32	10	1	10	0	1	32344M	15M	59M	96M	44M	
0	0	217	185	0	0	32	10	1	14	0	1	32344M	15M	59M	96M	44M	

The background is a dark teal gradient. It features several abstract elements: a large, semi-transparent teal circle on the left with a white dot at its center and a white line extending from it towards the top-left; a smaller teal circle at the top center with a white dot and a line extending towards the top-right; a cluster of two overlapping teal circles on the right with two white dots and lines extending towards the right edge; and a small teal circle at the bottom center with a white dot and a line extending towards the bottom-right.

tuned

Tuned

- What is tuned...pronounced “tune-D”
- Why was it created...
 - Why you should care ?
- Tuned in Red Hat products
- Tuned Profile Overview
- Improve User Experience for your App

Tuned Profiles in RH Product Line

RHEL7 Desktop/Workstation

balanced

RHEL7 Server/HPC

throughput-performance

RHEL6/7 KVM Host, Guest

Virtual-host/guest

RHEV

virtual-host

Red Hat Storage

rhs-high-throughput, virt

RHEL OSP (compute node)

virtual-host

RHEL Atomic

atomic-host, atomic-guest

Your Project

Next!

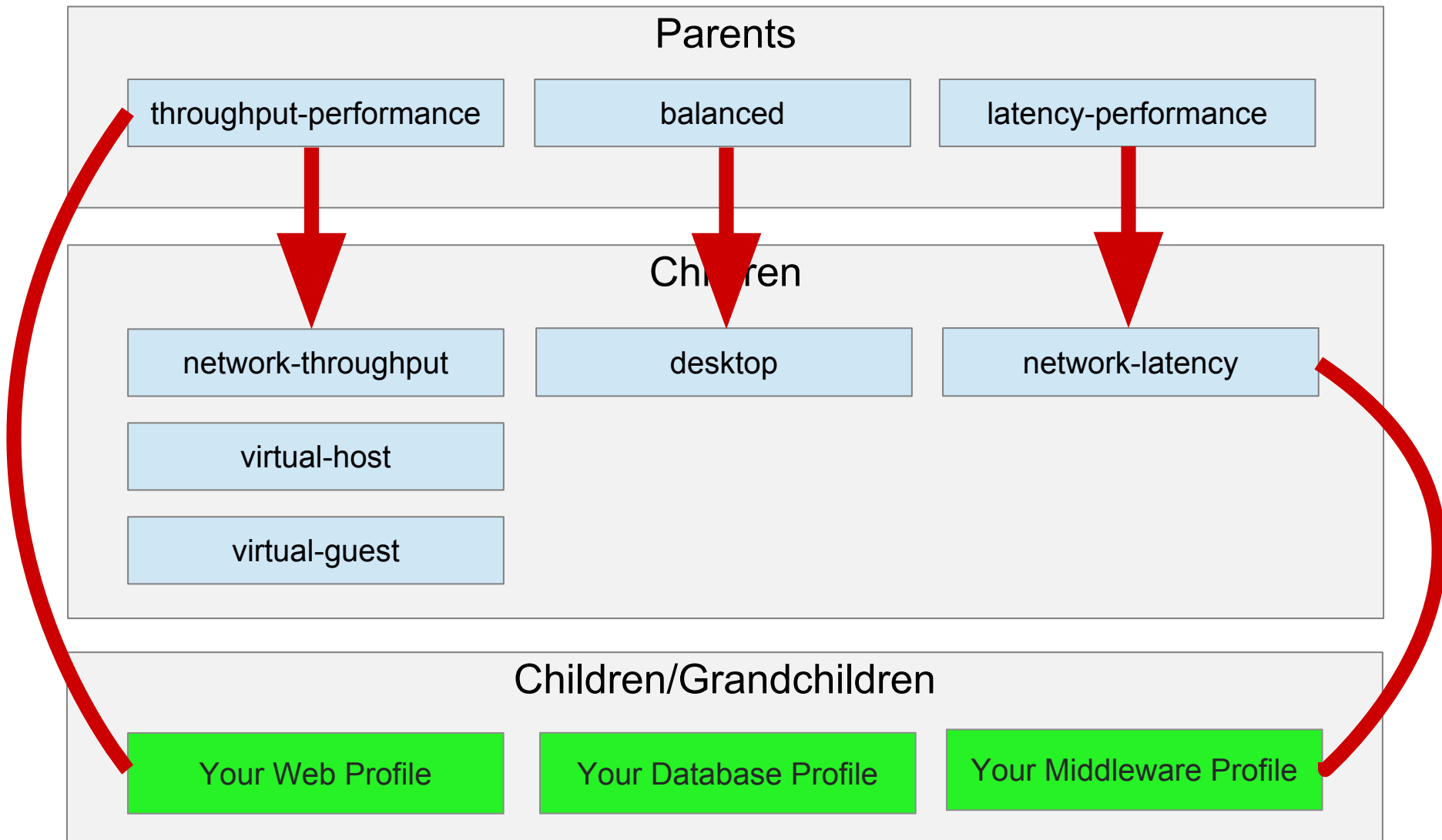
How do I create my own tuned profile on RHEL7 ?

<https://access.redhat.com/site/node/731473>

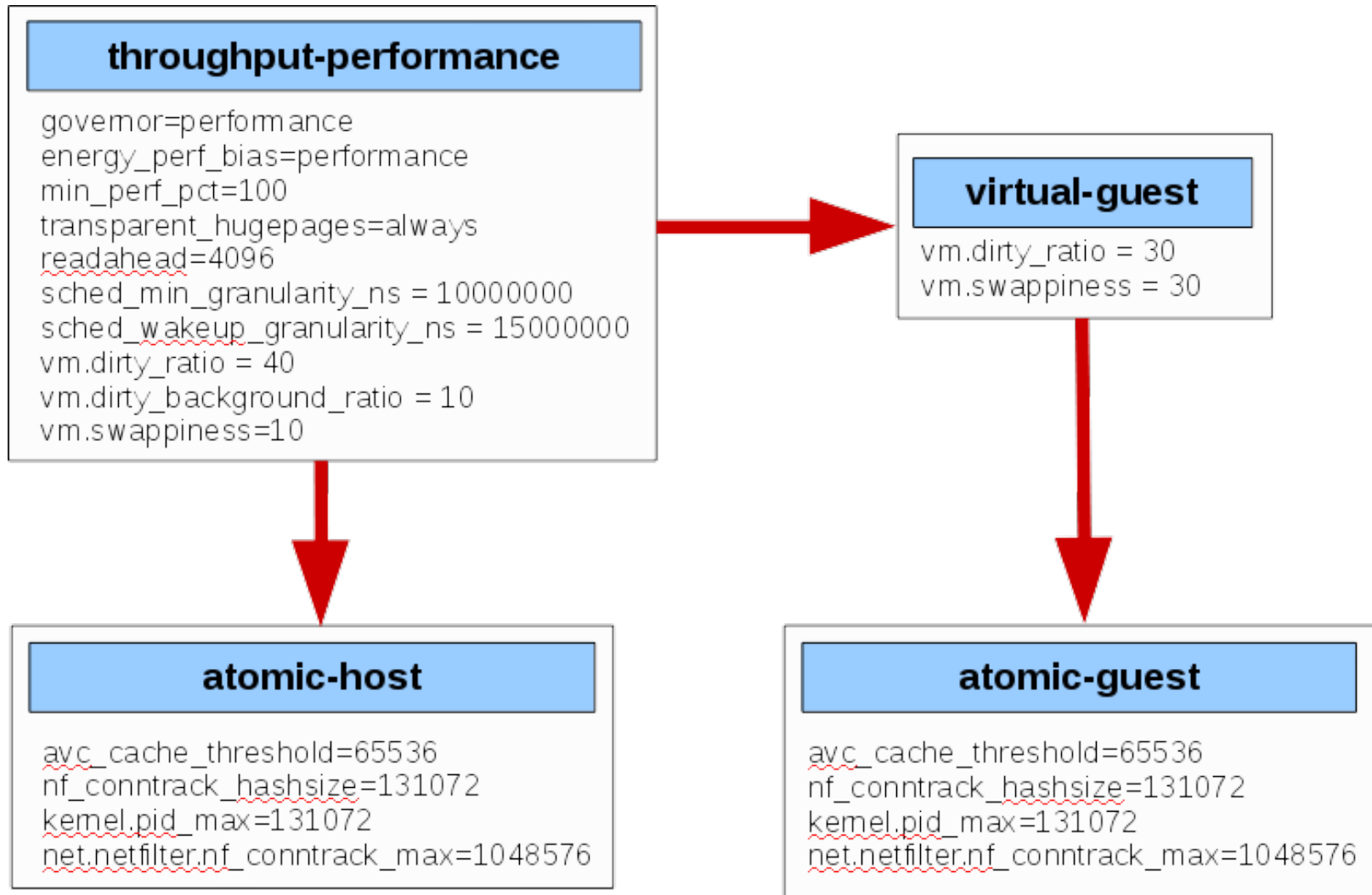
Tuned: Updates for RHEL7

- Installed by default!
- Profiles automatically set based on variant (recommend.conf)
- Optional hook/callout capability
- Concept of Inheritance (just like httpd.conf)
- Profiles updated for RHEL7 features and characteristics
- Added bash-completion :-)

Tuned: Your Custom Profiles



Tuned Profile Inheritance



Overview of Performance Analysis Utilities

perf

perf

Userspace tool to read CPU
counters and kernel tracepoints

perf list

- List counters/tracepoints available on your system

```
# perf list
```

```
List of pre-defined events (to be used in -e):
```

cpu-cycles OR cycles	[Hardware event]
instructions	[Hardware event]
cache-references	[Hardware event]
cache-misses	[Hardware event]
branch-instructions OR branches	[Hardware event]
branch-misses	[Hardware event]
cpu-clock	[Software event]
task-clock	[Software event]
page-faults OR faults	[Software event]
context-switches OR cs	[Software event]
cpu-migrations OR migrations	[Software event]
minor-faults	[Software event]
major-faults	[Software event]

perf list

- grep for something interesting, maybe to see what numabalance is doing ?

```
# perf list | grep sched: | grep numa
sched:sched_move_numa [Tracepoint event]
sched:sched_stick_numa [Tracepoint event]
sched:sched_swap_numa [Tracepoint event]
```

perf top

- System-wide 'top' view of busy functions

```
Samples: 10K of event 'cycles', Event count (approx.): 5973713325
34.35%    httpd [kernel.kallsyms] [k] avtab_search_node
12.70%    httpd [kernel.kallsyms] [k] _spin_lock
 8.61%    httpd [kernel.kallsyms] [k] tg_load_down
 7.42%    httpd [kernel.kallsyms] [k] _spin_lock_irq
 5.79%     init [kernel.kallsyms] [k] intel_idle
 3.92%    httpd [kernel.kallsyms] [k] _spin_lock_irqsave
 1.75%    httpd [kernel.kallsyms] [k] sidtab_search_core
 1.74%    httpd [kernel.kallsyms] [k] load_balance_fair
 1.18%    httpd [kernel.kallsyms] [k] tg_nop
 1.13%     init [kernel.kallsyms] [k] _spin_lock
```



perf record

- Record system-wide (-a)
- A single command
- An existing process (-p)
- Add call-chain recording (-g)
- Only specific events (-e)

```
# perf record -g dd if=/dev/zero of=1G bs=1M count=10 oflag=direct
10+0 records in
10+0 records out
10485760 bytes (10 MB) copied, 0.0830041 s, 126 MB/s
[ perf record: Woken up 1 times to write data ]
[ perf record: Captured and wrote 0.016 MB perf.data (~715 samples) ]
```

perf report

#	Overhead	Command	Shared	Object
#				
#				
	43.53%	dd	[kernel.kallsyms]	[k] __clear_user
		---	__clear_user	/dev/zero
			--99.75%--	read_zero.part.5
				read_zero
				vfs_read
				sys_read
				system_call_fastpath
				__GI___libc_read
			--0.25%--	[...]
	5.37%	dd	[kernel.kallsyms]	[k] do_blockdev_direct_IO
		---	do_blockdev_direct_IO	
			__blockdev_direct_IO	
			xfs_vm_direct_IO	
			generic_file_direct_write	
			xfs_file_dio_aio_write	
			xfs_file_aio_write	
			do_sync_write	

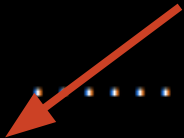
perf report

#	Overhead	Command	Shared Object
#			
#			
	43.53%	dd	[kernel.kallsyms] [k] __clear_user
		--- __clear_user	/dev/zero
		--99.75%-- read_zero.part.5	
			read_zero
			vfs_read
			sys_read
			system_call_fastpath
			__GI___libc_read
		--0.25%-- [...]	
	5.37%	dd	[kernel.kallsyms] [k] do_blockdev_direct_IO
		--- do_blockdev_direct_IO	oflag=direct
		__blockdev_direct_IO	
		xfs_vm_direct_IO	
		generic_file_direct_write	
		xfs_file_dio_aio_write	
		xfs_file_aio_write	
		do_sync_write	

perf diff

- Compare 2 perf recordings

```
# perf diff
# Event 'cycles'
#
# Baseline      Delta      Shared Object      Symbol
# .....
#
12.88% -12.27% [kernel.kallsyms] [k] __lookup_mnt
11.97% -11.17% systemd [.] 0x000000000000064968
4.32% +6.43% libdbus-1.so.3.7.4 [.] 0x000000000000029258
4.06% +4.72% dbus-daemon [.] 0x000000000000014a6e
3.79% -3.79% libglib-2.0.so.0.3600.3 [.] 0x000000000000088d6a
3.72% +0.25% [kernel.kallsyms] [k] seq_list_start
```



Helpful Utilities

Supportability

- redhat-support-tool
- sos
- kdump
- perf
- psmisc
- strace
- sysstat
- systemtap
- trace-cmd
- util-linux-ng

NUMA

- hwloc
- Intel PCM
- numactl
- numad
- numatop (01.org)

Power/Tuning

- cpupowerutils (R6)
- kernel-tools (R7)
- powertop
- tuna
- tuned

Networking

- dropwatch
- ethtool
- netsniff-ng (EPEL6)
- tcpdump
- wireshark/tshark

Storage

- blktrace
- iostat
- iostat

Questions