The kernel report

(LinuxCon Japan 2013 edition)

Jonathan Corbet
LWN.net
corbet@lwn.net

The kernel report (and weather forecast)

(LinuxCon Japan 2013 edition)

Jonathan Corbet
LWN.net
corbet@lwn.net

In the dim and distant past

In the dim and distant past

(May 20, 2012)

The 3.4 kernel was released

Since then...

69,863 changesets merged
3,173 developers have contributed
360+ employers have contributed
5 kernels released
The kernel is 1.57 million lines bigger

Recent release history

Release	Date	Days	Csets	Devs
3.4	May 20	63	10,899	1,286
3.5	July 21	62	10,957	1,195
3.6	Sep 30	71	10,247	1,216
3.7	Dec 10	71	11,990	1,271
3.8	Feb 18	70	12,394	1,253
3.9	Apr 28	69	11,910	1,339
3.10			12,776*	1,237*

* so far

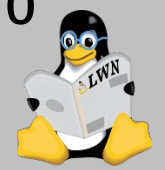
A few headline features

Bufferbloat fixes
Android integration
dm-verity, dm-cache
x32 ABI
Seccomp filters
User namespaces
uprobes
TCP fast open
Btrfs send/receive, RAID
Kernel hardening

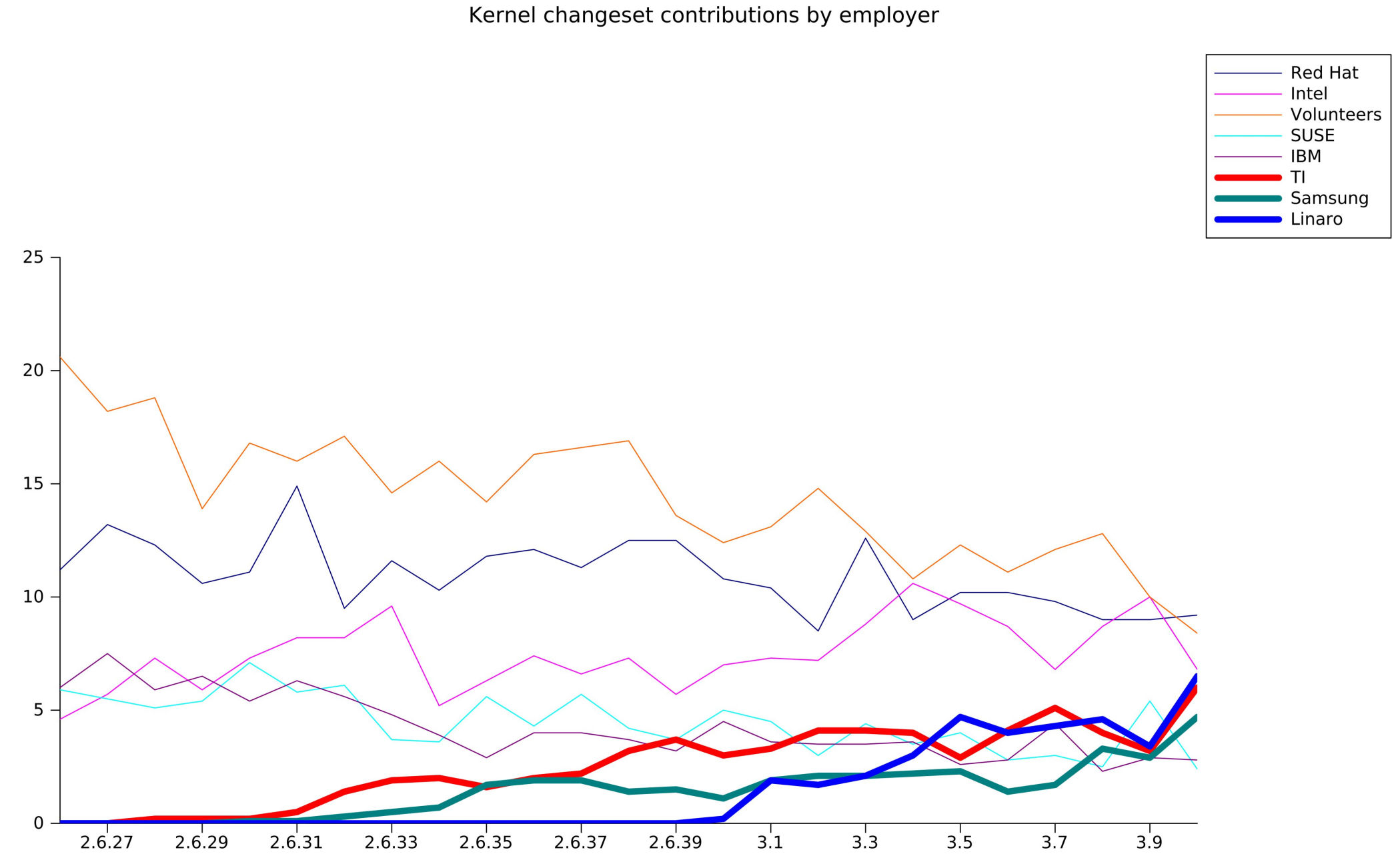
64-bit ARM support
OpenVSwitch
NUMA balancing
Kernel mem. acc'ting
ARM single zImage
ARM virtualization
Per-entity load tracking
Memory pressure notif.
Dynamic tick
XFS metadata checksum
...

Most active employers (3.4..)

(None)	11.3%	NVIDIA	1.4%
Red Hat	9.8%	Wolfson Micro	1.3%
Intel	8.4%	Oracle	1.3%
(unknown)	7.2%	Renasas Elect.	1.2%
Linaro	4.5%	Freescale	1.2%
Texas Inst.	4.2%	consultants	1.2%
SUSE	3.3%	Broadcom	1.2%
IBM	3.0%	Cisco	1.1%
Vision Engraving	3.0%	LINBIT	1.0%
Samsung	2.8%	Inktank Storage	1.0%
Google	2.5%	Ingics Tech	1.0%



Mobile/embedded participation



Lots going on out there

Lots going on out there
(as usual)

The 3.10 kernel

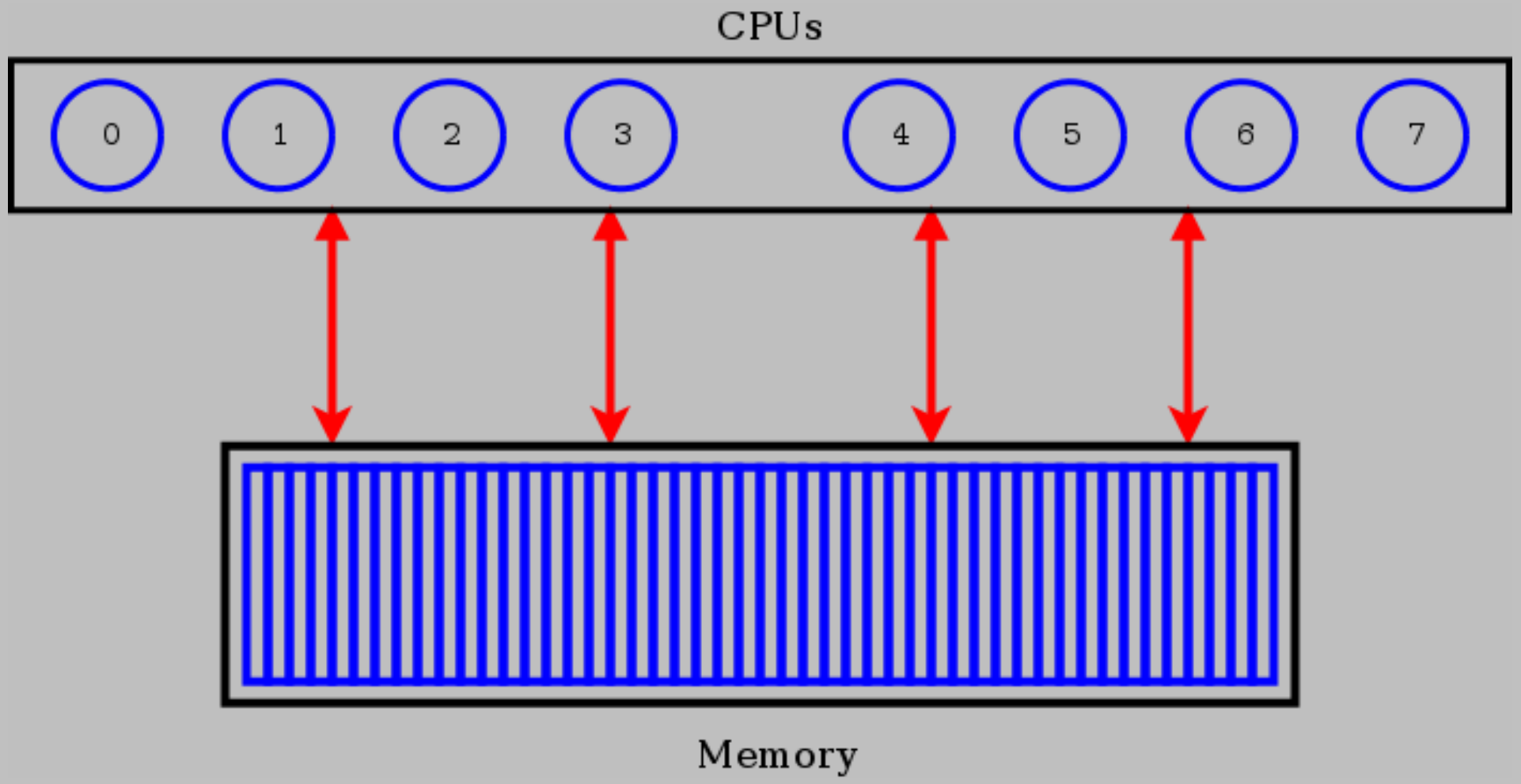
Due in early July

Features

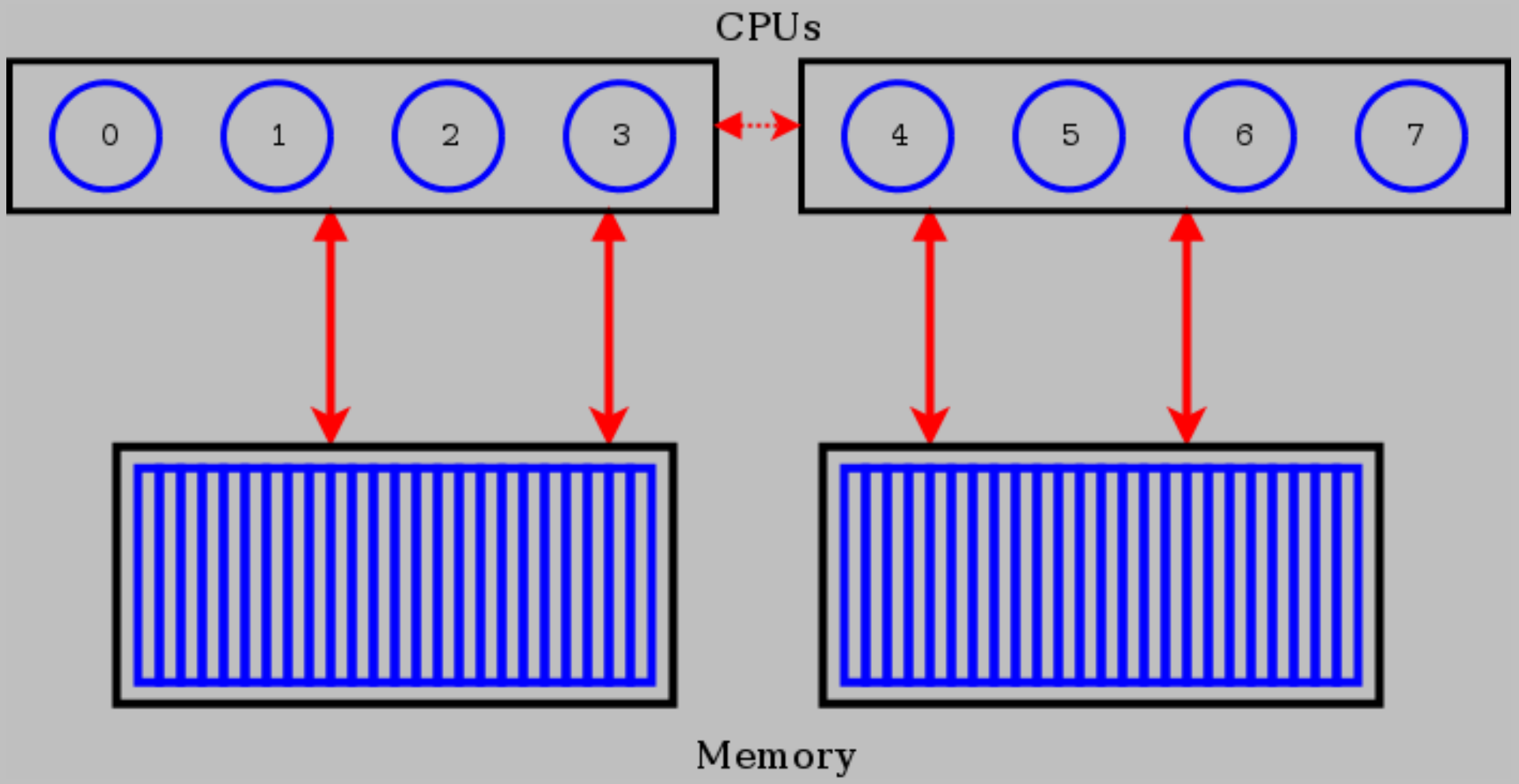
- (Nearly) full dynamic tick
- Bcache
- Multiple ftrace buffers
- Memory pressure notifications
- TCP tail loss probing
- ARM single zImage work
- XFS metadata checksums
- Arm big.LITTLE preparation
- ...

Scheduling

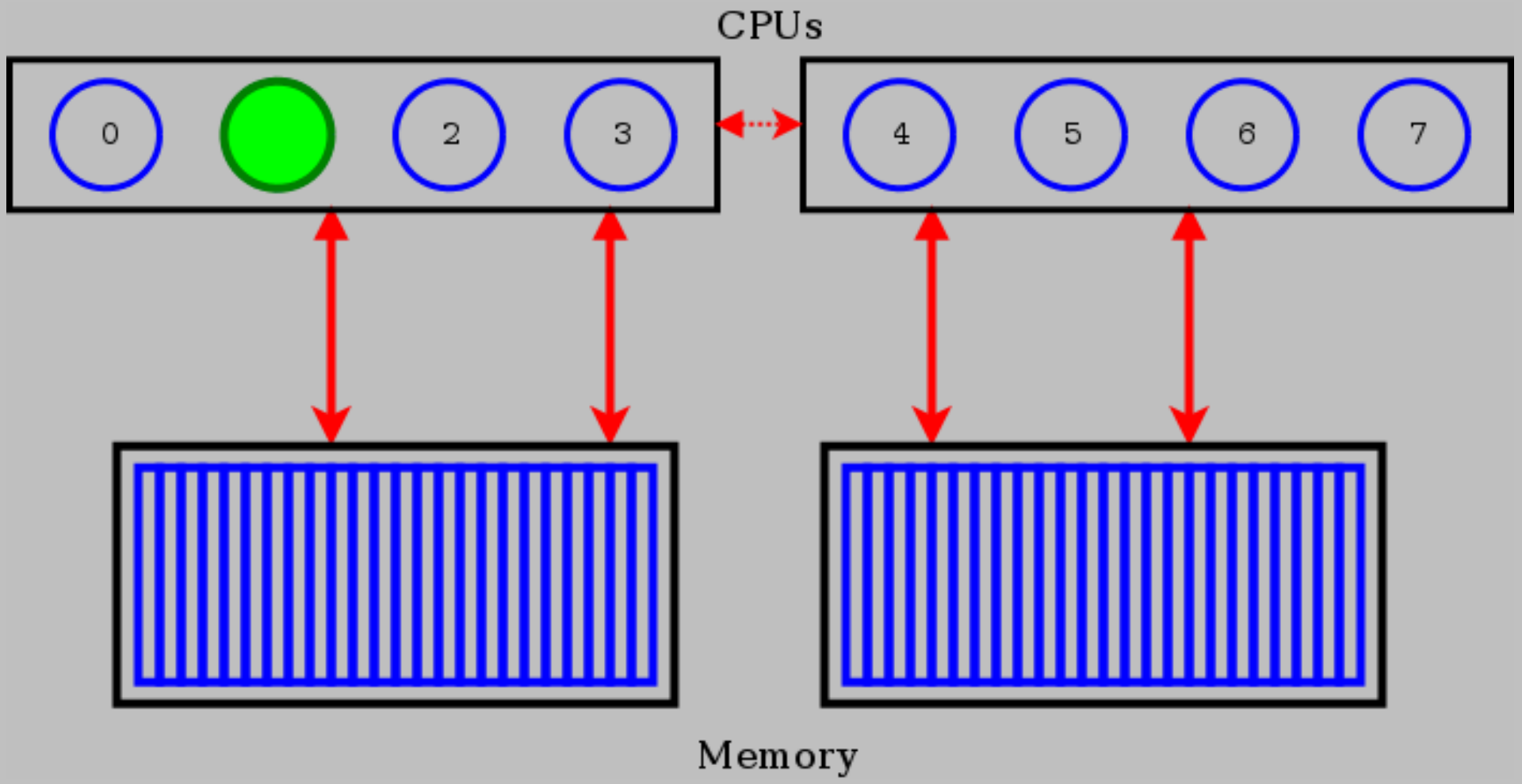
NUMA balancing



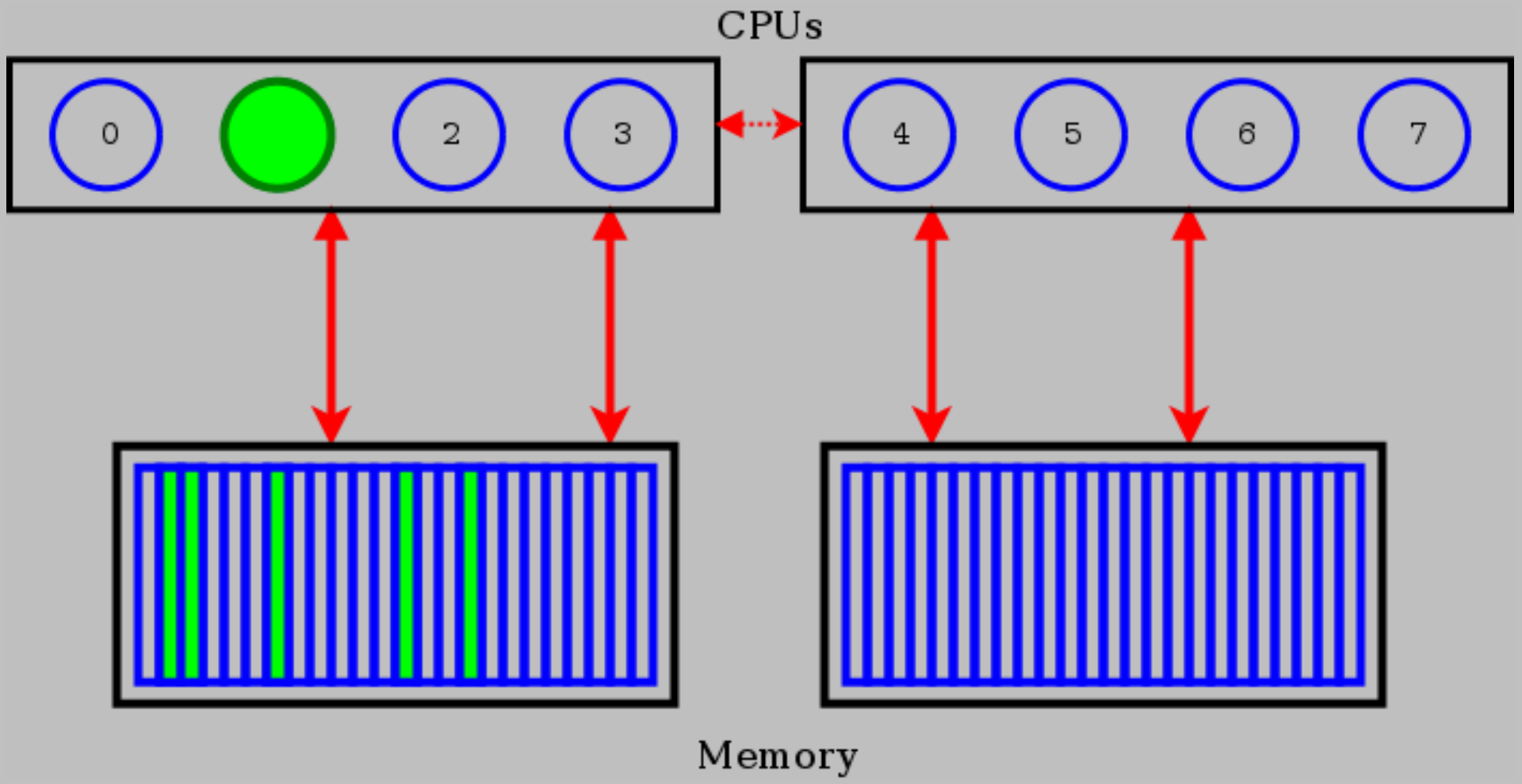
NUMA balancing



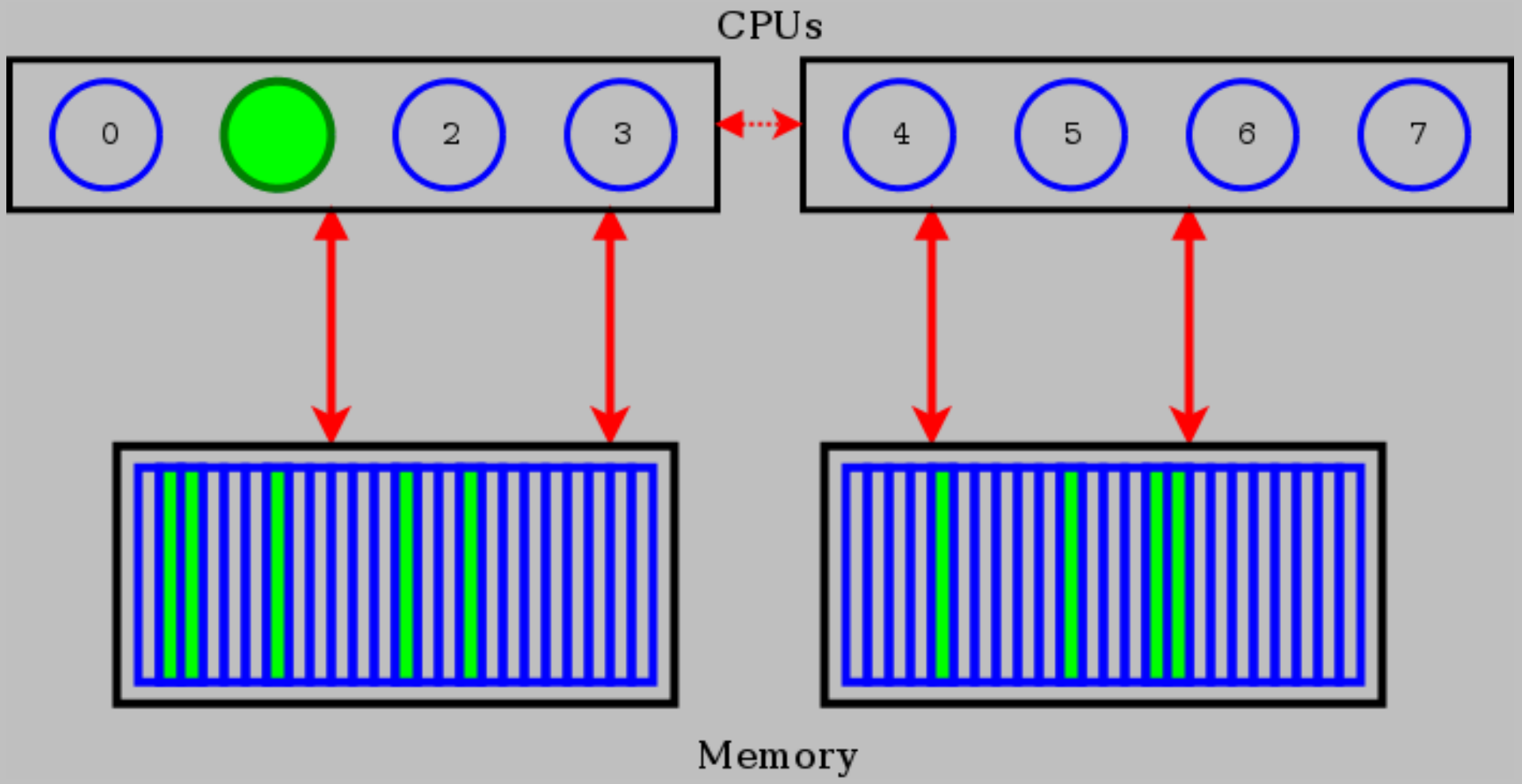
NUMA balancing



NUMA balancing



NUMA balancing



Toward better NUMA performance

NUMA scheduling framework added v3.8

Ideas:

- A “home node” for processes
- Actively migrate pages when processes move
- ...

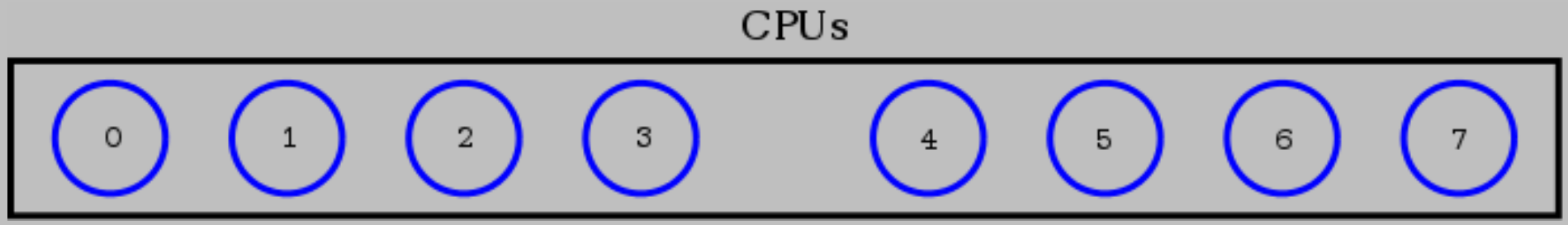
There's more

Power-aware scheduling

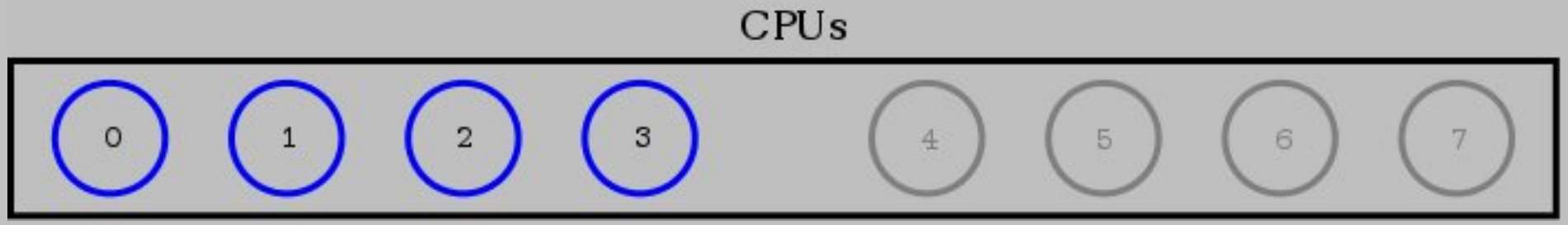
Linux is very good when there's no work to do

The problem is easy on a busy system

Power-aware scheduling



Power-aware scheduling



Approaches to power-aware scheduling

Pack small tasks together

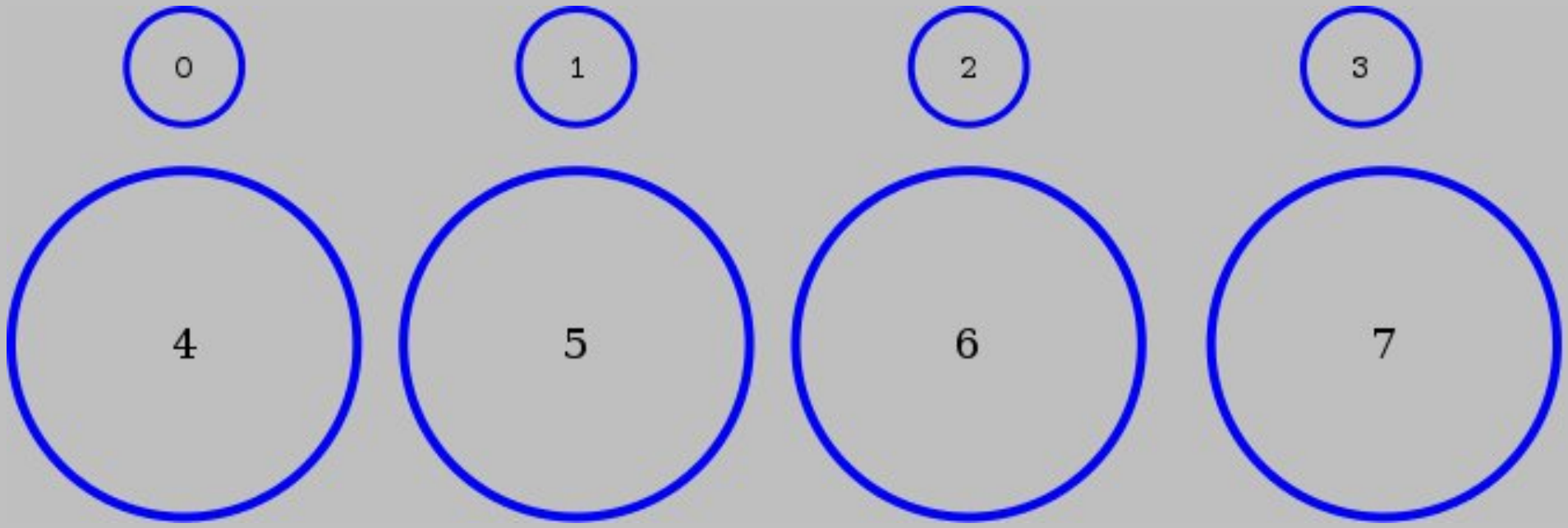
- Don't let them wake multiple processors

Spread big tasks out

- Try to get to idle quickly

Wait...there's more!

big.LITTLE



Approaches to big.LITTLE

Switch all CPUs transparently in a hypervisor

- An apparent 4-CPU system
- Kernel is unaware of switching

Approaches to big.LITTLE

Switch all CPUs transparently in a hypervisor
- An apparent 4-CPU system
- Kernel is unaware of switching

Bundle pairs of big and little CPUs
- Treat it like frequency scaling
- “The big.LITTLE switcher”

Approaches to big.LITTLE

Switch all CPUs transparently in a hypervisor
- An apparent 4-CPU system
- Kernel is unaware of switching

Bundle pairs of big and little CPUs
- Treat it like frequency scaling
- “The big.LITTLE switcher”

Teach the scheduler about heterogeneous systems
- “big LITTLE MP” patches exist
- The best solution someday

Where are the patches?

big.LITTLE switcher
- Not upstream
- Code ~~not yet~~ recently posted

big LITTLE MP
- Early-stage patches posted
- Trouble getting review and integration

Why is this stuff hard to merge?

Fear of regressions

Who hacks arch/arm* (3.3-3.9)

Linaro	15%
Texas Instruments	12%
(none)	7%
consultants	7%
ARM	6%
NVIDIA	6%
Samsung	5%
Renesas Electronics	3%
Atomide	3%
Free Electrons	3%
Freescale	3%
Calxeda	2%

Who hacks {kernel,mm,fs}/

Red Hat	25%
SUSE	8%
NetApp	8%
Google	5%
Parallels	5%
(none)	4%
IBM	4%
Oracle	4%
Linaro	4%
Fujitsu	4%
Intel	3%
Arista Networks	3%

“Linux kernel development is driven by the needs of enterprise computing.”

We have an interesting integration problem

Why bother?

Maybe we need two kernels?

- One for enterprise / desktop
- One for mobile / embedded

Features inflicted upon embedded by the enterprise

Symmetric multi-processing
Support for >1GB of memory on 32-bit systems
The ext3/ext4 filesystems
The completely fair scheduler
Best-of-breed networking
...

One kernel for everybody is one of our strengths

What about Android?

Much code merged

Some has been replaced
- Including wakelocks!

Some remains outside
- ION memory manager
- …

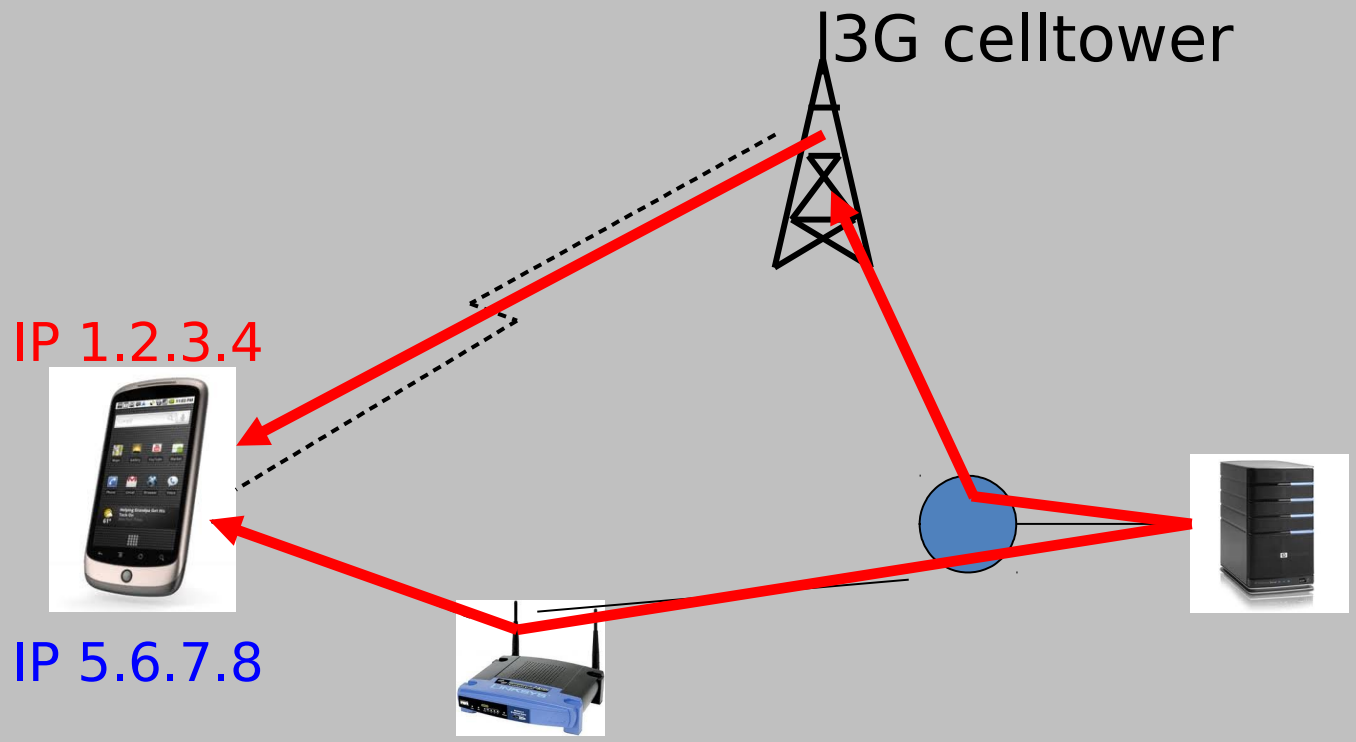
Google continues to innovate

Networking

Multipath TCP

Your phone has (at least) two interfaces

- Only one is used at a time



Multipath TCP

What if both could be used together?

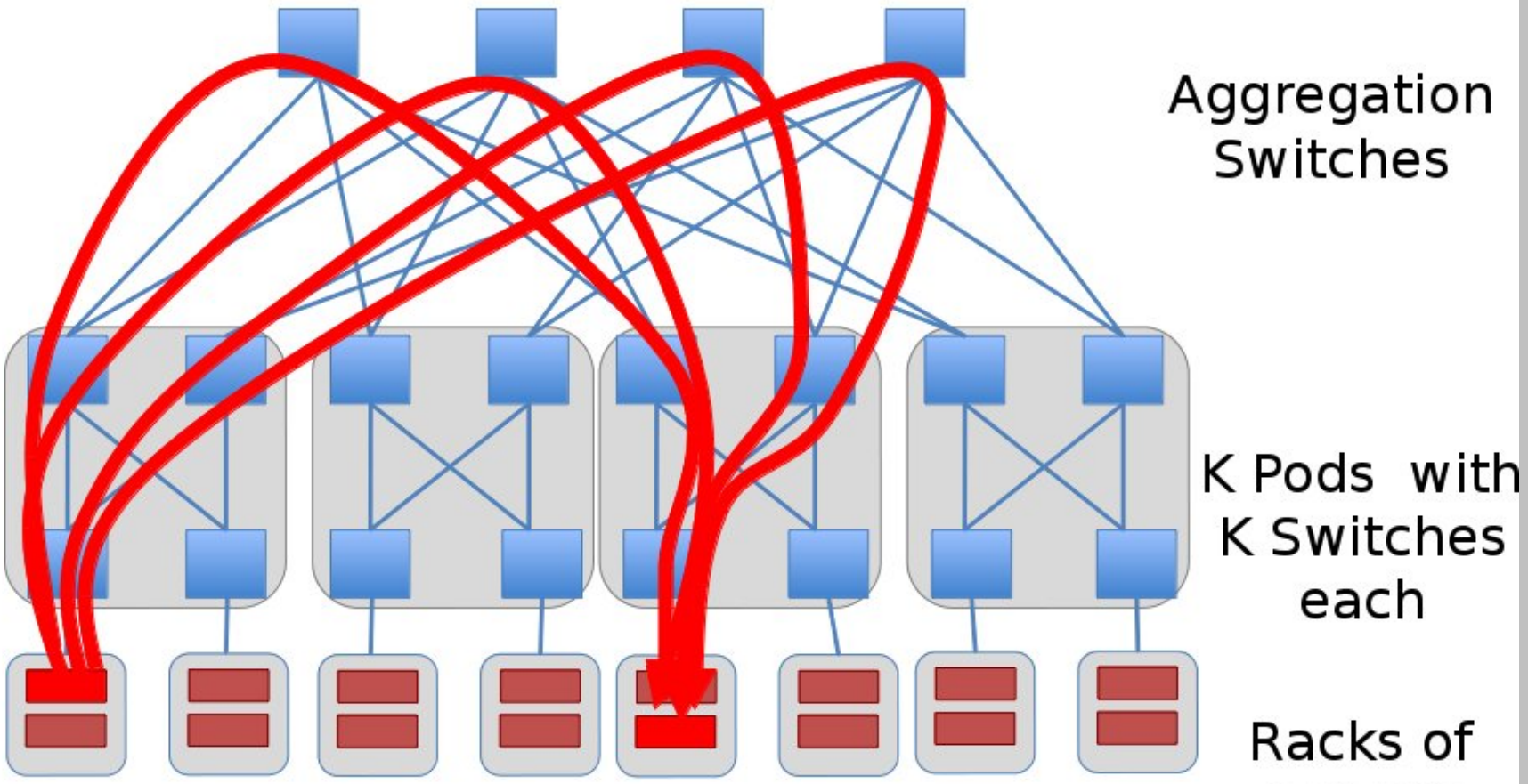
Better bandwidth

- Traffic can run across both links simultaneously

More reliable

- Paths can come and go - connection remains

Useful in data centers too



C. Raiciu, et al. “Improving datacenter performance and robustness with multipath TCP,” *ACM SIGCOMM* 2011.

Multipath TCP status

Patches exist

- Lots of information on multipath-tcp.org
- Speed record for TCP claimed
- Mainline merging still distant

But that's not the interesting part...

The innovation-hostile net

Multipath TCP is complex

- Flow management
- Congestion control
- Security issues
- …

Where the real challenge was:

- The dreaded middlebox

Middleboxes

These boxes will:

- Change IP addresses on the fly (ports too)
- Add or remove TCP flags
- Resegment data
- ACK undelivered data
- Corrupt the data itself
- Block anything that looks “scary”

Thus:

- MPTCP must look **exactly** like TCP

Network innovation comes to Linux first

Network innovation comes to Linux first

...but...

Deploying new protocols on the net has become nearly impossible.

Other interesting stuff

CPU isolation

When you must have the whole CPU to yourself

With 3.10:

- No timer tick
- No RCU callbacks
- ...

Useful for:

- High-performance computing
- Realtime

Security

Attackers are reviewing our code better than we are

Some improvements

Better testing
- Trinity - fuzz testing

Kernel hardening efforts

Useful new technologies
- User namespaces
- ...

It's still not good enough.

Innovation

Different views of Linux

- Classic Linux
  - Kernel
  - GNU libc
  - SYSV init
  - X11
  - Desktop env

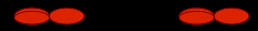
Different views of Linux

Classic Linux	Android
Kernel	Kernel
GNU libc	Bionic
SYSV init	init.rc
X11	SurfaceFlinger
Desktop env	Dalvik

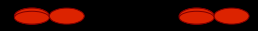
Different views of Linux

Classic Linux	Android	Ubuntu
Kernel	Kernel	Kernel
GNU libc	Bionic	GNU libc
SYSV init	init.rc	upstart
X11	SurfaceFlinger	Mir
Desktop env	Dalvik	Unity

Are we repeating the Unix wars?



Interesting things are now done on Linux first



It's up to us to find the path(s)