

The Kernel Report

(Linux-Kongress 2008 edition)

Jonathan Corbet
LWN.net
corbet@lwn.net

Tracking 2.5

Jonathan Corbet
corbet@lwn.net

This talk is available at:

- <http://lwn.net/talks/linux-kongress/>

Theme

Challenges / Responses

Challenge

Get the next release out

Response

The 2.6.x release cycle

4-5 releases per year
Each a major release

2.6.23 – October 9, 2007

CFS scheduler

First mac80211 driver

Xen core

Lguest

2.6.24 – January 24, 2008

Network namespaces

Control groups

i386/x86_64 architecture merger

Kernel markers

2.6.25 – April 16, 2008

ath5k wireless driver

SMACK security module

Video driver updates (R500)

Realtime group scheduling

ext4 filesystem improvements

memory usage controller

2.6.26 – July 13, 2008

x86 PAT support

Read-only bind mounts

More network namespace work

KGDB

2.6.27 – any day now

Block layer data integrity checking

Ftrace

gspca video camera drivers

UBIFS

Multiqueue networking

System call extensions – new flags

Challenge

Sustain a high rate of development
One of the fastest anywhere

A single kernel cycle involves
10,000+ individual changesets
1,000 developers
1-200 corporations

A single kernel cycle involves
10,000+ individual changesets
1,000 developers
1-200 corporations

2.6.27:
10,600 changesets
1109 developers
150 companies

linux-next

Contains patches for 2.6.n+1

Find integration problems

Early testing

The new development kernel

...sort of

Challenge

Maintaining kernel quality

Too many features, too few fixes?

Responses

Tracking and fixing of regressions

Listed regressions statistics:
















Date	Total	Pending	Unresolved

2008-09-12	163	51	38
2008-09-07	150	43	33
2008-08-30	135	48	36
2008-08-23	122	48	40
2008-08-16	103	47	37
2008-08-10	80	52	31
2008-08-02	47	31	20

Responses

Better tools

4011 BUG/BUG_ON's reported

firegl_ioctl		2139	[external] Bug in the proprietary fireglx driver
page_remove_rmap		299	
set_page_address		255	
remove_wait_queue		135	
klist_add_tail		85	Bug in the w9968cf_usb driver
default_idle		59	
drm_open		59	
__mutex_lock_slowpath		50	
rt2500pci_config_intf		46	
__ieee80211_if_config		43	
exit_mmap		43	
cpu_idle		39	
acpi_idle_enter_bm		32	
mwait_idle		30	
ipw_send_cmd		30	

Responses

Social pressure + tighter rules

“Here's a simple rule of thumb:
if it's not on the regression list
if it's not a reported security hole
if it's not on the reported oopses list
then why are people sending it to me?”
-- Linus Torvalds

Challenge

The kernel is a common resource
...driven by divergent interests

Response

The “upstream first” policy
No differentiation at the kernel level

Who contributes

2.6.23 -> 2.6.27-rc5

(None)	19%	Movial	2%
Red Hat	12%	SGI	1%
IBM	7%	academia	1%
unknown	6%	Analog Devices	1%
Novell	6%	Renasas Tech	1%
Intel	5%	Freescala	1%
Parallels	2%	MontaVista	1%
Oracle	2%	Fujitsu	1%
linutronix	2%	Google	1%
consultants	2%	Astaro	1%

Challenge

Out-of-tree code

Challenge

Out-of-tree code

Binary-only modules

Vendor-private code

External projects

Responses

Developer outreach

Merging outside projects
Even if the code isn't great

Discouraging binary modules

Challenge

Security

Challenge

Security
...of the kernel itself

Challenge

Security

- ...of the kernel itself

- ...support for user-space security

2008 CVEs (Jan - August)

CVE-2008-3792 CVE-2008-3686 CVE-2008-3535
CVE-2008-3534 CVE-2008-3526 CVE-2008-3525
CVE-2008-3496 CVE-2008-3276 CVE-2008-3275
CVE-2008-3272 CVE-2008-3247 CVE-2008-3077
CVE-2008-2931 CVE-2008-2826 CVE-2008-2812
CVE-2008-2750 CVE-2008-2729 CVE-2008-2372
CVE-2008-2365 CVE-2008-2358 CVE-2008-2148
CVE-2008-2137 CVE-2008-2136 CVE-2008-1675
CVE-2008-1673 CVE-2008-1669 CVE-2008-1619
CVE-2008-1615 CVE-2008-1375 CVE-2008-1367
CVE-2008-1294 CVE-2008-0600 CVE-2008-0598
CVE-2008-0352 CVE-2008-0010 CVE-2008-0009
CVE-2008-0007 CVE-2008-0001

2008 CVEs (Jan - August)

CVE-2008-3792 CVE-2008-3686 CVE-2008-3535
CVE-2008-3534 CVE-2008-3526 CVE-2008-3525
CVE-2008-3496 CVE-2008-3276 CVE-2008-3275
CVE-2008-3272 CVE-2008-3247 CVE-2008-3077
CVE-2008-2931 CVE-2008-2826 CVE-2008-2812
CVE-2008-2750 CVE-2008-2729 CVE-2008-2372
CVE-2008-2365 CVE-2008-2358 CVE-2008-2148
CVE-2008-2137 CVE-2008-2136 CVE-2008-1675
CVE-2008-1673 CVE-2008-1669 CVE-2008-1619
CVE-2008-1615 CVE-2008-1375 CVE-2008-1367
CVE-2008-1294 **CVE-2008-0600** CVE-2008-0598
CVE-2008-0352 CVE-2008-0010 CVE-2008-0009
CVE-2008-0007 CVE-2008-0001

Responses...?

User-space security

Unix-style DAC may not be enough

User-space security

SELinux

SMACK

AppArmor

TOMOYO Linux

TALPA / fanotify

Challenge

Scalability

Scalability issues

Locking

Contention kills performance
Cache effects hurt

Solutions

Finer-grained locking
Lockless algorithms



Scalability issues

Memory use

Scalability issues

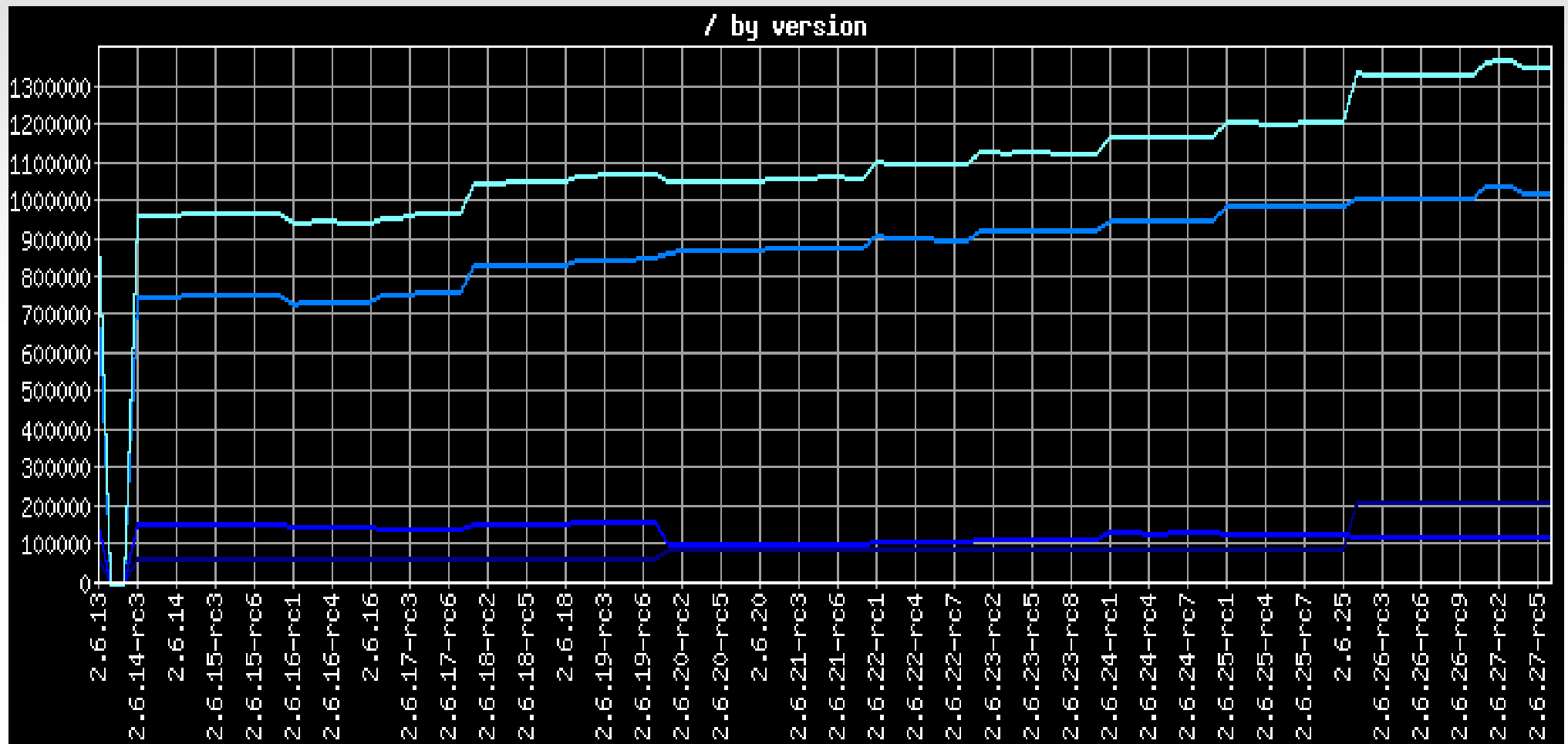
Memory use



Solution: better data structures

Scalability goes both ways

Scalability issues



Scalability issues

What to do?

- Pay attention to bloat

- Small-system configuration options

- More participation from embedded folks

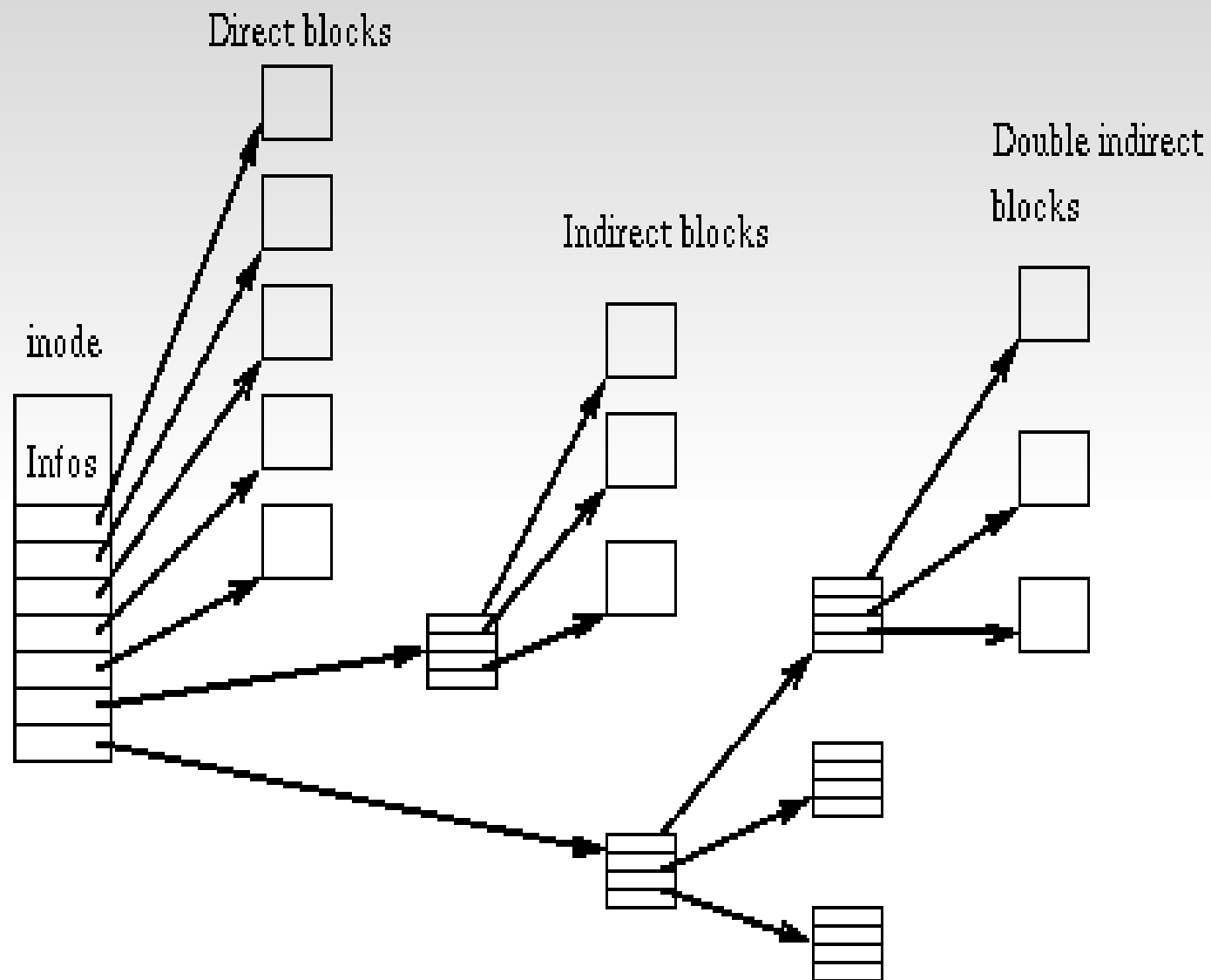
Challenge

Storage and filesystems

Disks were small
...as were files

(DEC RP06, 178 MB)





A 4GB DVD image

1M (4K) disk blocks

- 12 direct blocks

- 1024 via the single-indirect block

- ~1M via the double-indirect block

A 4GB DVD image

1M (4K) disk blocks

- 12 direct blocks

- 1024 via the single-indirect block

- ~1M via the double-indirect block

There must be a better way

Other problems

- fsck takes forever

- Limits on file and filesystem sizes

- No data integrity protection

- No snapshots

- Generally old

Response: ext4

The progression of ext3

- Extents

- Better allocation

- File and filesystem limits lifted

- Journal checksums

Response: btrfs

A completely new filesystem

Extents

Subvolumes

Snapshots

Full checksumming

Fast fsck

Challenge

Solid-state storage

Truly random-access

Fast reads, slow writes

Wear leveling required

Our current flash filesystems

...are showing their age

Responses

Btrfs

UBIFS

Merged for 2.6.27

Expects direct access to flash

Logfs

Seemingly stalled

Challenge

Hardware support

Responses

Life just gets better

AMD/ATI releases information

Atheros hires community developers

VIA employs a community liaison

Sometimes life improves slowly

Wireless networking

Video adapters

Help life get better yet

- Avoid closed hardware

- Avoid binary-only drivers

- Avoid uncooperative companies

Challenge

Power management
Better battery life

Challenge

Power management

Better battery life

Better planetary life

Responses

Lots of work across the board

Better drivers

Better core support

Better user space

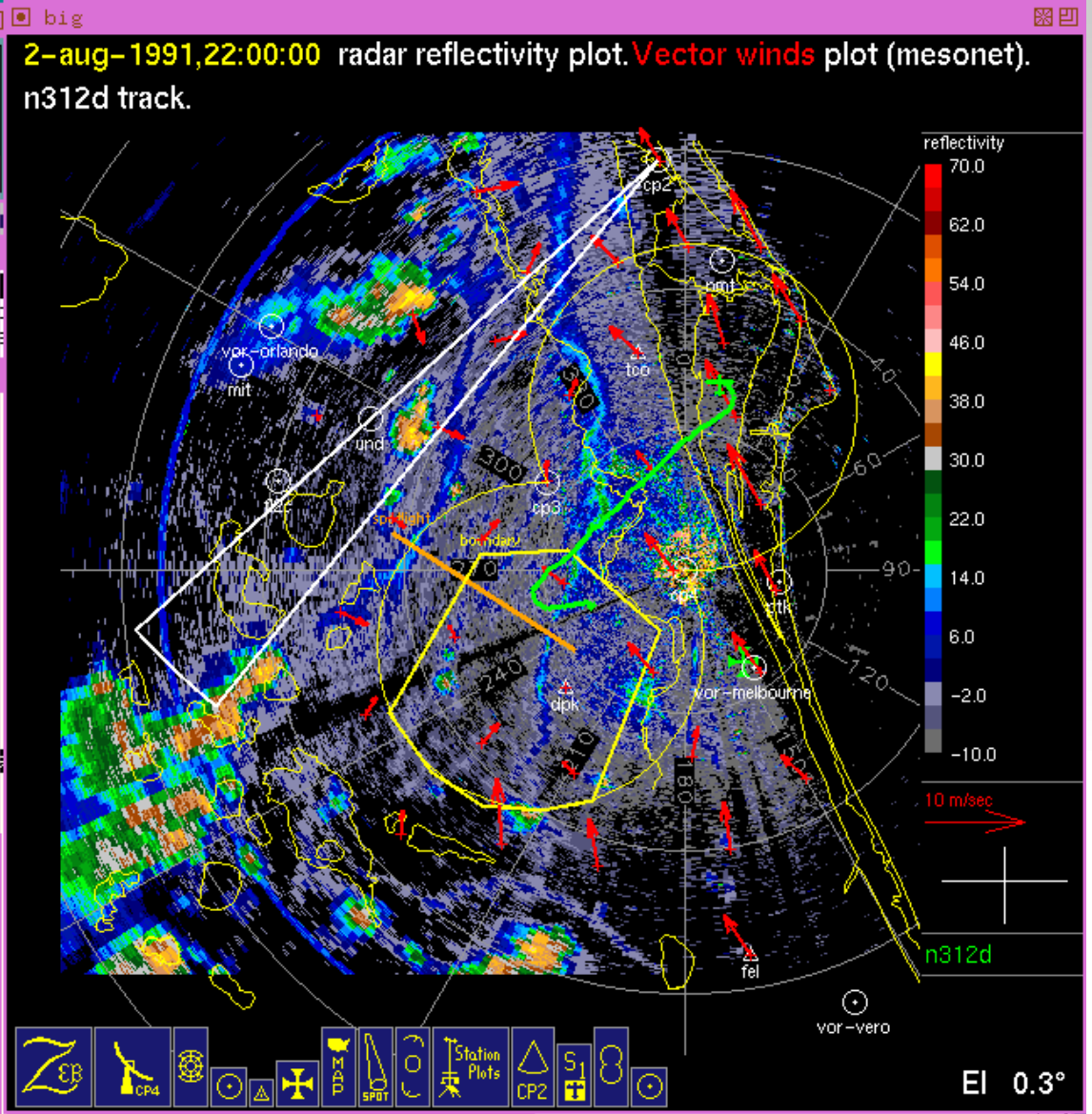
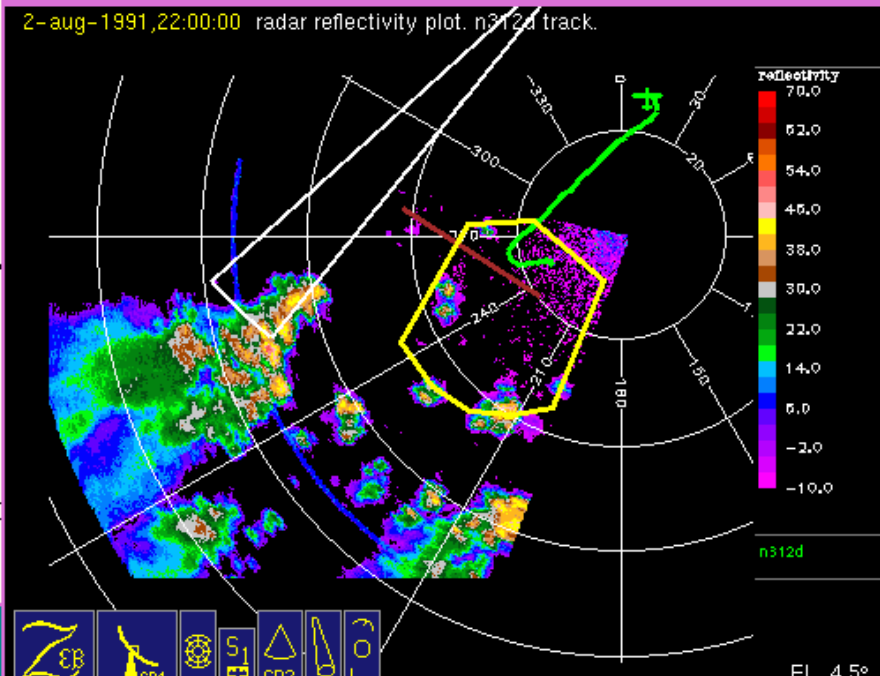
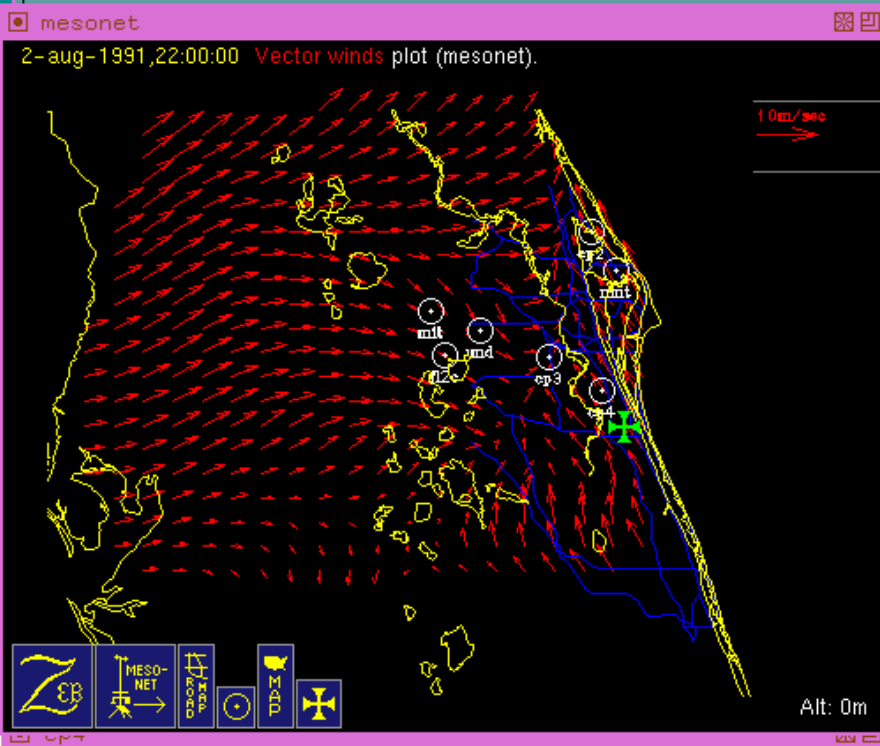
Challenge

Hard real-time support

Who needs realtime?

Data acquisition / process control

iconbar



position

Get Position

Lat: 28 11'21" N Lon: 81 8'6" W

Az: 263 Range: 21 Nm

x: -39.2 km y: -4.5 km

Relative to cp4

Deg/Min/Sec

Nn

movie

Movie Control:

Movie for minutes, every minutes, ending at Frames/Second:

Status:

Who needs realtime?

Commercial exchanges



Who needs realtime?

Gadgets



Realtime responses

Realtime group scheduling
Stabilizing in 2.6.27

The realtime patch set
Sleeping spinlocks
Threaded interrupt handlers
Lots of other stuff

Challenge

Maintain the best network stack

Responses

Lots of internal work
Wireless networking
IPv6 improvements
Network channels

...

Challenge

Virtualization

Responses

Xen

Still improving

KVM

Where the action seems to be

Lguest

Puppy-safe virtualization

Challenge

Containers



(photo: Darin Marshall)

Responses

Much code already merged

- Control groups

- Resource controllers

- Network, PID, user, ... namespaces

Some still waiting

- Sysfs support

- Checkpoint and restore

- Management support

Challenge

Tracing

Responses

SystemTap

- Powerful tool

- Dynamic tracing

- Painful to use

- No user-space tracing

Responses

Other tracing tools

ftrace

LTTng

Why not just port DTrace?

Questions?