Mixing kvm and xen: Xenner

or booting xen pv guest kernels in kvm

Gerd Hofffmann <kraxel@redhat.com>

Talk Outline

- xenner intro & design
- current state & future plans
- benchmark numbers

arch overview

- xenner -- userspace app.
- emu{32,64} -- runs within the vm.
- vmcore file -- backing store.
- libxenctrl + xen bits (xenstored, qemu-dm, pygrub).
- xenner daemons.
- xenner tools.

xenner, the application

- cmd line tool, qemu-like for easier libvirt integration.
- creates xenish start-of-day enviroment using libxenguest.
- handles xenstore & event channels.
- one thread per vcpu, one main thread (I/O, signals).
- uses libkvm, slightly modified.

emu{32,64} -- part 1

- runs inside the vmx/svm container
- uses “out $magicport” to vmexit into xenner.
- virtual memory layout:
  - like xen.
- physical memory layout.
  - 4 MB memory text+data for emu.
  - 4 MB p2m table (more for larger guests).
  - everything else is assigned to the guest, 1:1 mapping shifted by 8 MB.

emu{32,64} -- part 2

- initially emu was a bunch of trap entry points with “out $port” and xenner handled it.
- moved more and more code into emu
  - correctness: xenner changing page tables doesn't fly.
  - performance: fewer vmexits.
  - interfaces: reduce xenner <=> emu dependencies.
  - security/stability: fewer places where xenner handles guest data.
- almost everything is done by emu these days.

emu{32,64} -- part 3

- uses in-kernel lapic
  - timer
  - IPI
- uses in-kernel i/o apic
  - event channels
- uses paravirt clocksource
  - patches in the queue ...
  - still off by default

emu{32,64} -- part 4

- memory management
  - no page table verification
  - just turn on the write bit for kernel pages on faults.

vmcore -- backing store

- /var/run/xenner/vmcore.$domid
- located on tmpfs
- ELF core file.

libxenctrl

- shared library, drop-in replacement.
- implements part of the xen library interface.
  - open/close + a few more basic calls.
  - event channel interface.
  - grant table interface (gntdev).
  - overloads open() (fake /proc/xen/ files).
- xen daemons (xenstored, xenconsoled, qemu-dm) run unmodified with that library.

evtchnd

- event channel implementation
- runs as daemon
- libxenctrl redirects event channel calls to this daemon.

blkbackd

- block backend driver
- built on top of libxenctrl too.
- works also with Xen.
  - within the limits of the gntdev driver.
- handles raw images (using aio).
  - I'm using this with lvm volumes.
- handles qemu formats (via bdrv_{read,write}).

netbackd

- network backend driver, using libxenctrl.
- host connection via tap device.
- just passing ethernet frames.
- no advanced stuff yet (TSO, sg, ...).

xenner tools

- xenner-stats
  - extracts statistics from vmcore.
  - also prints kvm stats from debugfs.
- xenner-cleanup
  - cleans up after xenner crashing.
  - if you need that you've found a bug.

current state -- features

- 32 / 32pae / 64 guests.
- SMP support.
- blk / net / console /pvfb backends.
- some bugs which prevent some guests from installing successfully.

current state -- guests

- rhel-5: installs and runs, LTP runs ok.
- fedora-8: installs and runs.
- fedora-9:
  - 32bit: installs and runs.
  - 64bit: install oopses kvm-69.
- opensuse 10.3: installs and runs
  - tested 32bit only.

future plans

- Test more guests, fix bugs as they show up.
- Use more kvm paravirt features.
- Look at 64bit performance.
- Implement more features (which ones ?).
- support non-{vmx,svm} systems (?)
- net backend: needs tuning.
- blk backend: aio for qemu formats.
- need long term maintainance master plan.

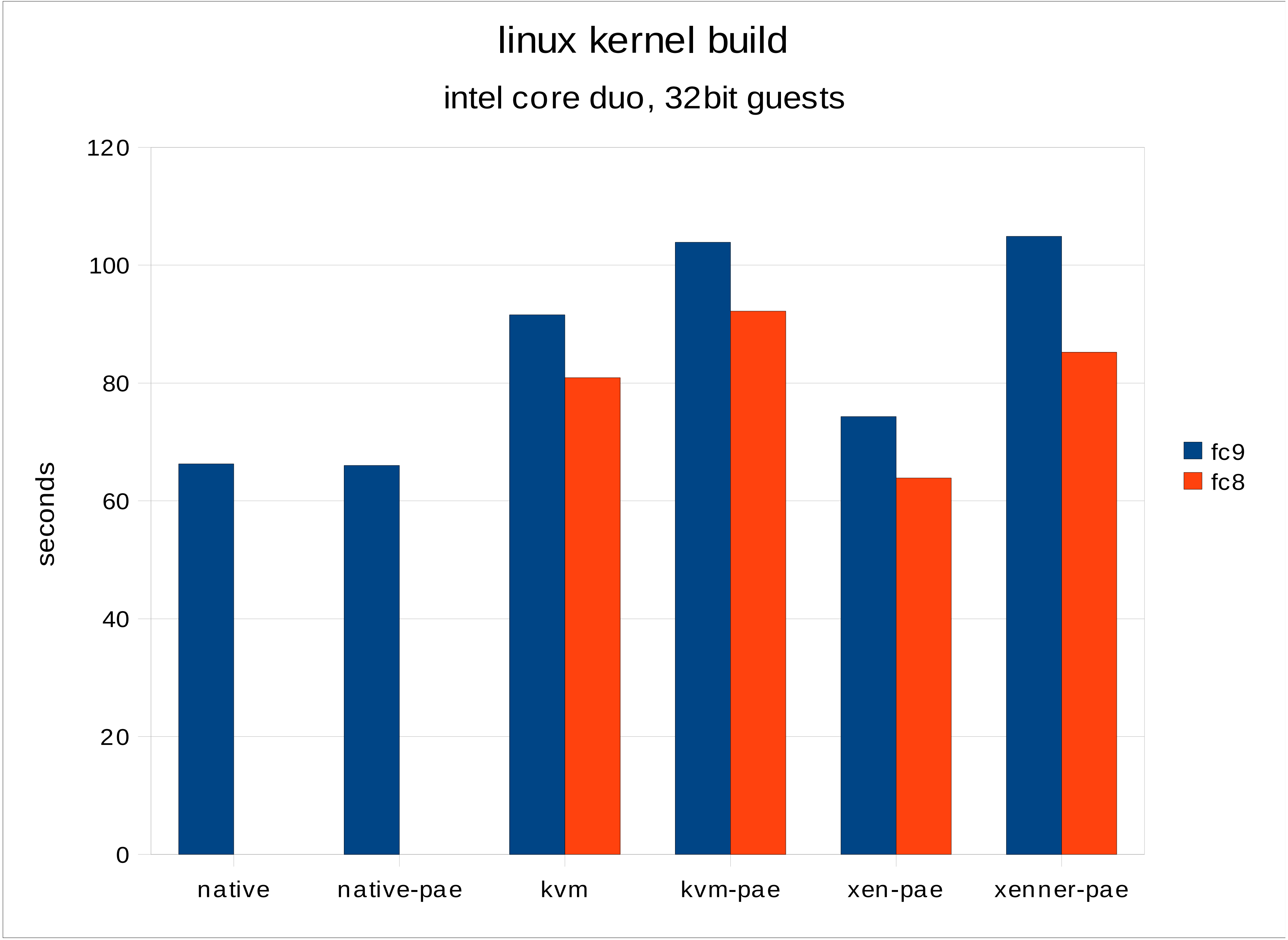
merging into qemu?

- Might make sense once I reached the point where the emu <=> xenner interfaces are stable.
- Problem: lost in a forest of qemu trees.
  - xen: qemu-dm for pvfb.
  - xenite: xen without xend, qemu-dm and libvirt doing the job instead.
  - kvm: qemu-kvm.
- Handle fullvirt guests with xen pv drivers?

lies, damn lies and benchmarks

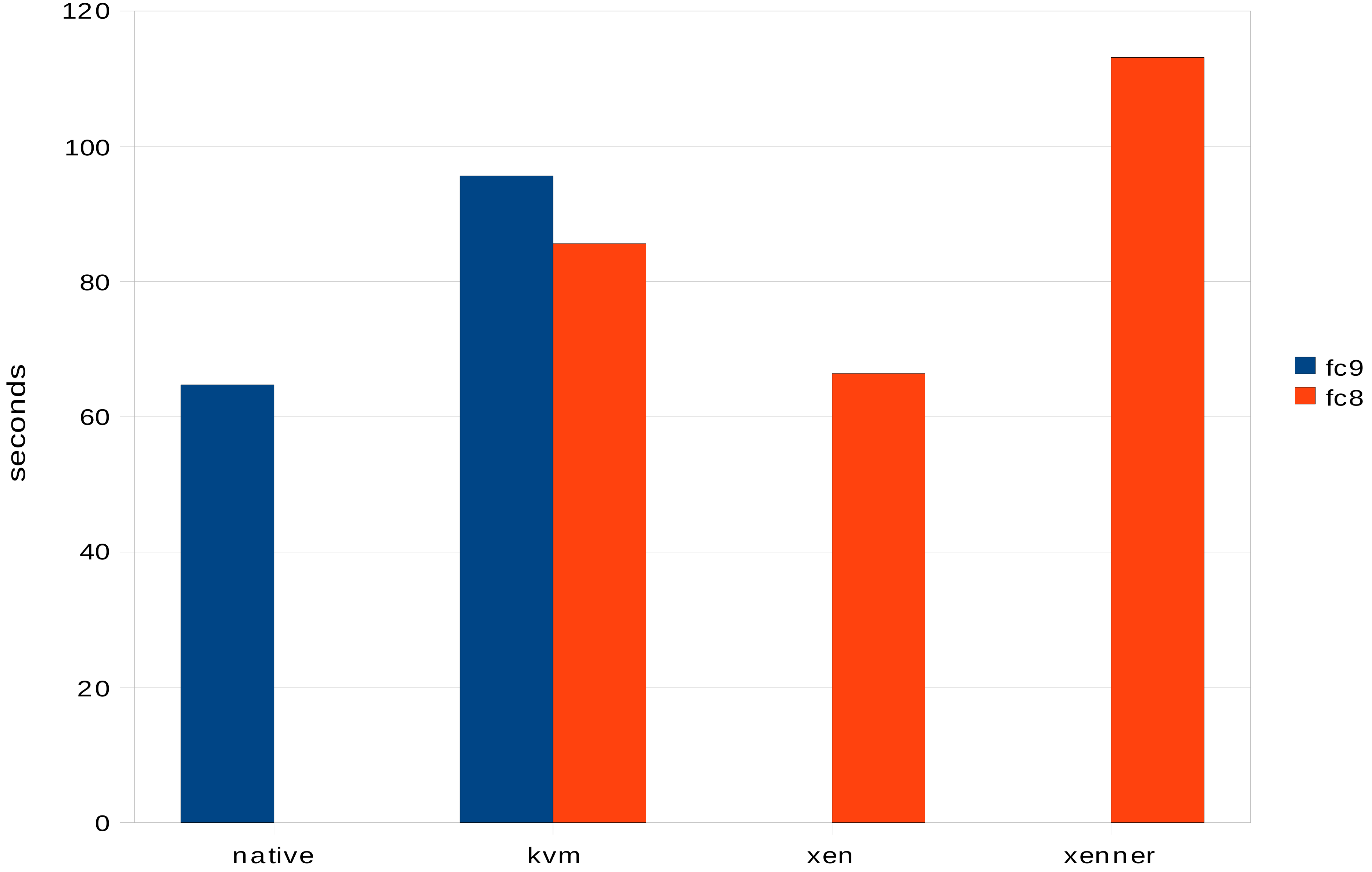
linux kernel build

- compiling: 2.6.25, i386, allnoconfig
- kvm host: Fedora 9, 64bit
  - 2.6.25
  - with kvm-69 modules for NPT tests.
- xen host: RHEL 5.2, 64bit
  - 2.6.18.
- guests:
  - Fedora 9 (2.6.25 pv_ops).
  - Fedora 8 (2.6.24, 2.6.21-xen).


linux kernel build
intel core duo, 32bit guests
seconds
120
100
80
60
40
20
0
native
native-pae
kvm
kvm-pae
xen-pae
xenner-pae
fc9
fc8

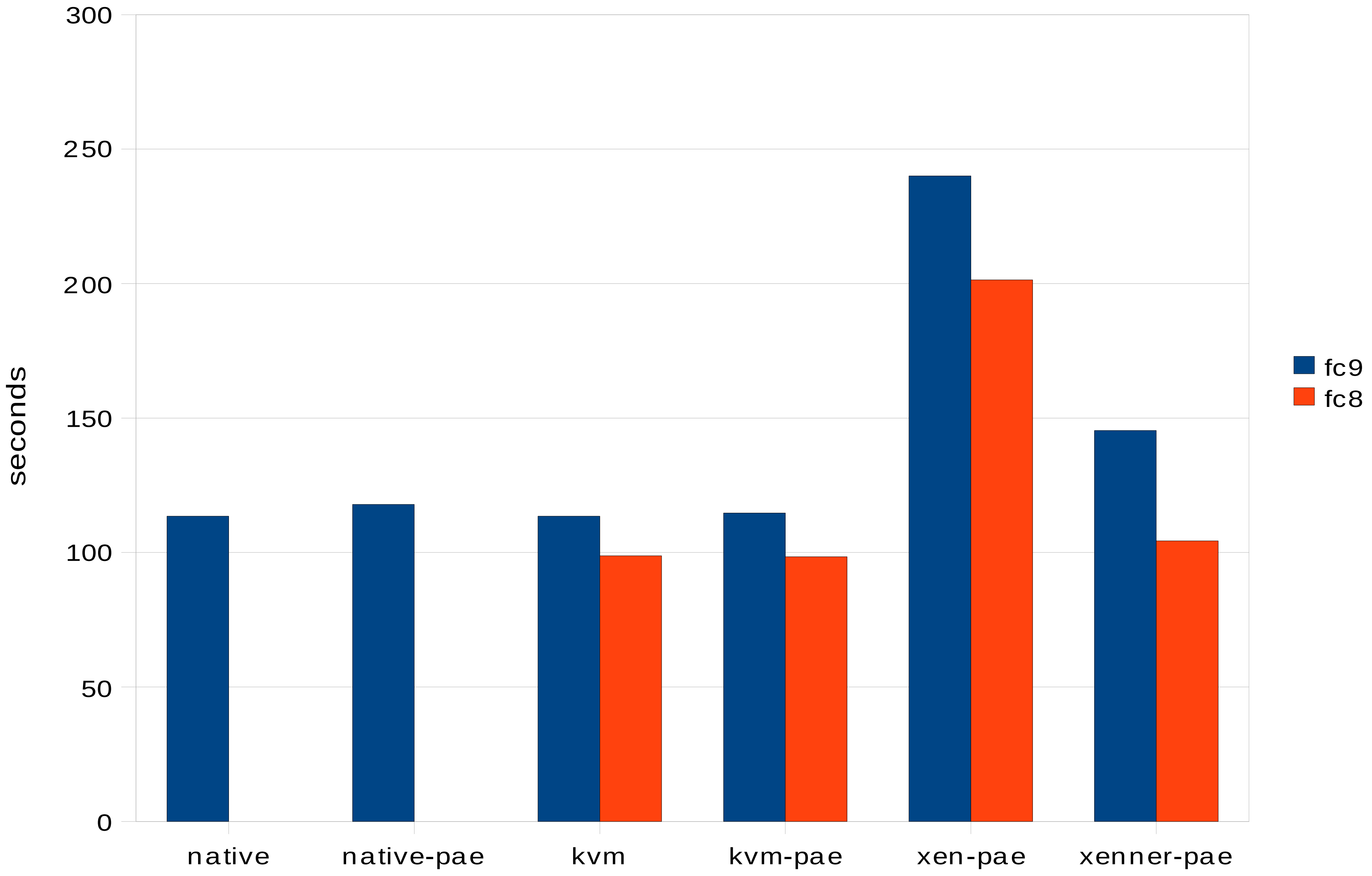
linux kernel build

intel core duo, 64 bit



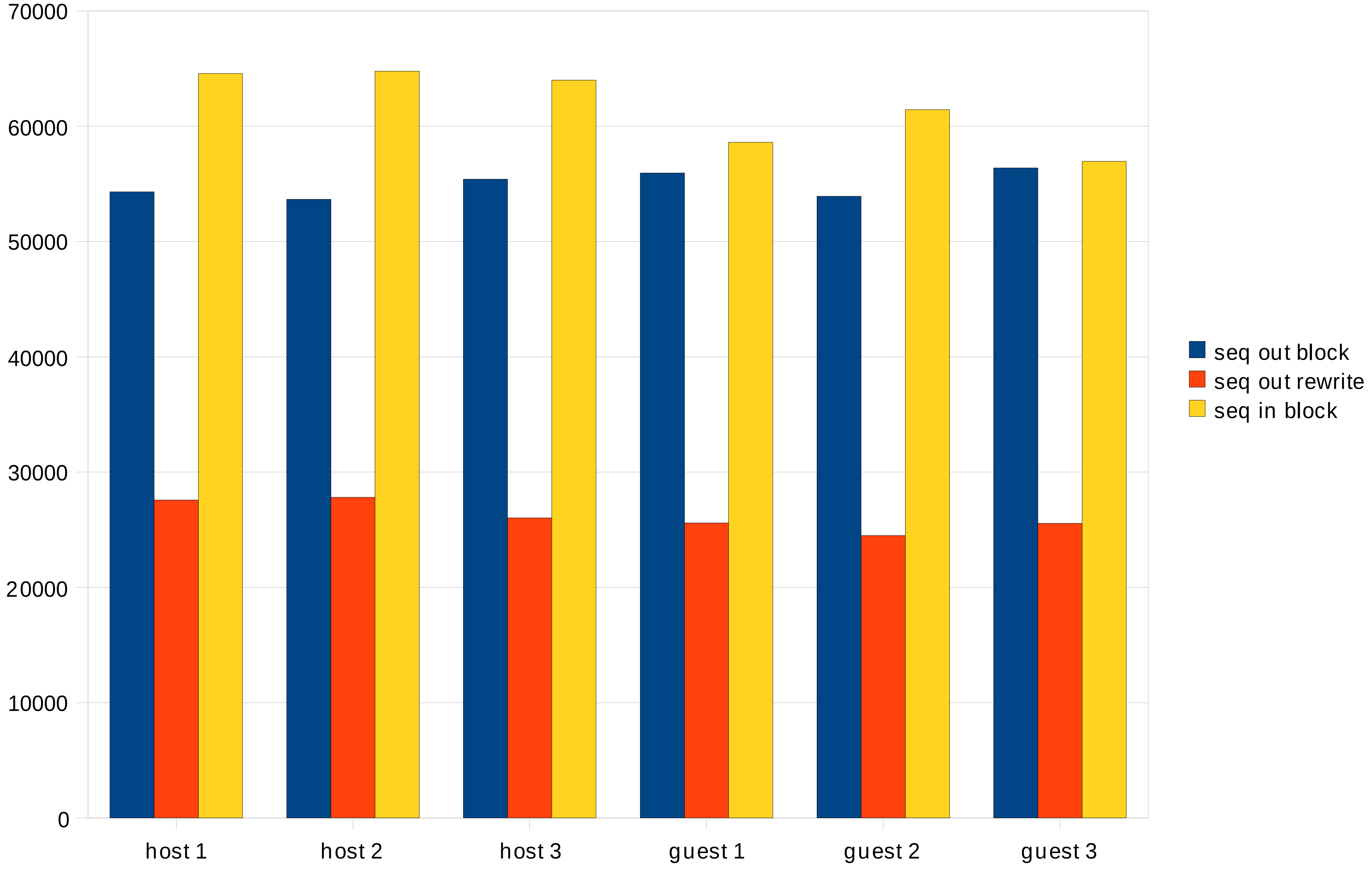
linux kernel build

amd barcelona, 32bit, NPT



bonnie++

512 MB RAM, 1G data set


host 1
host 2
host 3
guest 1
guest 2
guest 3
seq out block
seq out rewrite
seq in block