Dropping legacy devices in Qemu

Gabriel Laskar <gabriel@lse.epita.fr>

Why do we need a smaller VM?

- Reduced boot time
- Smaller attack surface
- Performances
- Why not?

ISA Devices

- On a fixed io address (non discoverable, no hotplug)
- Slow devices
- Under the Qemu Global lock for most of them
- MMIO are faster than IO in Qemu

What is necessary?

- Some kind of e820 support
- Devices (disk, nic): virtio devices
- Bus: PCIe bus should be enough
- CPUs and APICs: gathered through ACPI tables
- Timers & RTC: LAPIC, hpet & KVM PV clock
- Some way to load linux

Yep, no BIOS/EFI/Firmware.
Don’t need, don’t care.

Direct kernel boot

- Follow the linux boot protocol
- Skip the Real Mode kernel setup
- Directly boot into the PM mode code
- Feed setup with e820 tables, cmdline, initrd…

```
$ qemu-system-x86_64 -machine virt \
    --enable-kvm -serial stdio \
    --kernel $KBUILD_OUTPUT/arch/x86/boot/bzImage \
    -append "earlyprintk=serial,0x3f8,115200
console=ttyS0"
```

```
Decompressing Linux... Parsing ELF... Performing relocations... done.
Booting the kernel.
[    0.000000] Linux version 4.12.0-rc4+ (gaby@guinness) (gcc version 6.3.1 20170306 (GCC) )
#14 SMP Wed Jun 28 11:48:49 CEST 2017
[    0.000000] Command line: earlyprintk=serial,0x3f8,115200 console=ttyS0
[    0.000000] x86/fpu: x87 FPU will use FXSAVE
[    0.000000] e820: BIOS-provided physical RAM map:
[    0.000000] BIOS-e820: [mem 0x0000000000000000-0x0000000007ffffff] usable
[    0.000000] bootconsole [earlyser0] enabled
[    0.000000] NX (Execute Disable) protection: active
[    0.000000] DMI not present or invalid.
[    0.000000] Hypervisor detected: KVM
[    0.000000] tsc: Fast TSC calibration failed
[    0.000000] tsc: Unable to calibrate against PIT
[    0.000000] tsc: No reference (HPET/PMTIMER) available
[    0.000000] e820: last_pfn = 0x8000 max_arch_pfn = 0x400000000
[    0.000000] MTRR: Disabled
[    0.000000] x86/PAT: MTRRs disabled, skipping PAT initialization too.
[    0.000000] x86/PAT: Configuration [0-7]: WB  WT  UC- UC  WB  WT  UC- UC
[    0.000000] CPU MTRRs all blank - virtualized system.
Memory KASLR using RDTSC...
[    0.000000] Scanning 1 areas for low memory corruption
[    0.000000] ACPI: Early table checksum verification disabled
[    0.000000] ACPI BIOS Error (bug): A valid RSDP was not found (20170303/tbxfroot-244)
```

ACPI

- HW_REDUCED_ACPI flag in FADT
- Build tables and put them into RAM:
  - RSDP
  - RSDT
  - FADT
  - MCFG
  - HPET
  - DSDT (crippled, with only the bare minimum)
- Simple?

Qemu Bios_linker_loader

- Interface to load tables for firmware
- Must finish the tables link and checksums

Wait… pc-lite? What was that again?

- https://github.com/01org/qemu-lite
- http://events.linuxfoundation.org/sites/events/files/slides/Light%20weight%20virtualization%20with%20QEMU%26KVM_0.pdf