Western Digital®

RISC-V NOMMU and M-Mode Linux

Damien Le Moal, Western Digital

Linux Plumbers Conference, September 9th, 2019

Outline

Acknowledgement: Most of this work is being driven by Christoph Hellwig

- RISC-V NOMMU support: Why ?
- Ongoing work
  - Kernel side
  - User space and toolchains
- Demonstration
  - QEMU and Kendryte K210

RISC-V NOMMU support: Why ?

Not all about "because we can" ☺

- Support for CPUs lacking a memory management unit
  - mmu-type = "none" in device tree
    - sv39, sv48 define regular RISC-V MMU
  - No virtual memory, kernel and applications run in a single address space
  - Support for ARM, M68K, microblaze, SH, xtensa upstream
- This is not the same as user mode Linux !
  - Running on real hardware

- NOMMU and execution modes
  - S-mode is synonym with "have MMU"
    - NOMMU implies M-Mode
  - ECALLs from M-mode into M-mode are still possible
    - SBI can be used, but low value
  - M-mode support directly from the kernel a better fit for NOMMU
    - Direct IPI and timer control using MMIO (no SBI emulation)
    - Interrupt levels (S-mode and M-mode differ)
    - Support for user space behavior (signals/VDSO)

→ A lot of this work directly benefits regular S-mode (MMU) Linux by reducing SBI overhead with direct hardware support of performance sensitive functions (e.g. timer, IPI)

Ongoing Work: Kernel

V3 NOMMU series posted on 08/14

- Main changes
  - Introduce new config options
    - CONFIG_RISCV_M_MODE
    - CONFIG_RISCV_SBI
  - Abstract privileged CSRs names/numberscommon between S-mode and M-mode
    - E.g. sie/mie → xie
    - Definition controlled with CONFIG_RISCV_M_MODE
  - Completely disable SBI calls
  - Refactor and optimize IPI code
    - No SBI, use MMIO
    - Same for timer
  - Disable paging
  - Fix *rt_sigreturn* VDSO only implementation
- Allow bare metal boot
  - Introduce flat image
    - direct load at 0x80000000
  - No runtime FW
    - Tweaks head.S (HartID, trampoline)
  - New “nommu” defconfig file

QEMU NOMMU boot

```
qemu-system-riscv64 -smp 2 -m 64 -machine virt -nographic \
          -kernel arch/riscv/boot/loader \
          -drive file=rootfs.ext2,format=raw,id=hd0 \
          -device virtio-blk-device,drive=hd0
```

Ongoing Work: User space and toolchains

Flat bin only for now

- Relies on uClibc-ng (uCLinux) binfmt_flat format
  - No FDPIC loading support for now
    - Not supported by uClibc nor musl-libc
- More work needed for FDPIC support
  - Kernel 64-bit FDPIC loader support needed
    - Patch tested
  - Libraries initialization (loading)
    - Nothing done yet, as far as I know
- Resources
  - Kernel patches
    - http://git.infradead.org/users/hch/buildroot.git/shortlog/refs/heads/riscv-nommu.2
  - Buildroot image build
    - http://git.infradead.org/users/hch/riscv.git/shortlog/refs/heads/riscv-nommu.3

Ongoing Work: FDPIC Support

Some successes, need a *lot* more debug work

```
[    0.000000] Linux version 5.1.0-rc2-71673-ga4fabb1ca9dd (damien@washi) (gcc version 8.2.0 (Buildroot 2018.11-rc2-00003-ga0787e9)) #64 SMP Fri Apr 19 15:06:30 JST 2019
[    0.000000] Initial ramdisk at: 0x(____ptrval____) (271360 bytes)
[    0.000000] Zone ranges:
[    0.000000]   DMA32    [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000]   Normal   empty
[    0.000000] Movable zone start for each node
[    0.000000] Early memory node ranges
[    0.000000]   node   0: [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000] Initmem setup node 0 [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000] elf_hwcap is 0x112d
[    0.000000] percpu: max_distance=0x18000 too large for vmalloc space 0x0
[    0.000000] percpu: Embedded 12 pages/cpu @(____ptrval____) s18016 r0 d31136 u49152
[    0.000000] Built 1 zonelists, mobility grouping off.  Total pages: 2020
[    0.000000] Kernel command line:
[    0.000000] Dentry cache hash table entries: 1024 (order: 1, 8192 bytes)
[    0.000000] Inode-cache hash table entries: 512 (order: 0, 4096 bytes)
[    0.000000] Sorting __ex_table...
[    0.000000] Memory: 6276K/8192K available (947K kernel code, 101K rwdata, 168K rodata, 85K init, 97K bss, 1916K reserved, 0K cma-reserved)
...
[    0.283408] clocksource: Switched to clocksource riscv_clocksource
[    0.357702] Unpacking initramfs...
[    0.541675] Freeing initrd memory: 264K
[    0.567316] workingset: timestamp_bits=62 max_order=11 bucket_order=0
[    0.723861] Serial: 8250/16550 driver, 1 ports, IRQ sharing disabled
[    0.757731] 10000000.uart: ttyS0 at MMIO 0x10000000 (irq = 10, base_baud = 230400) is a 16550A
[    0.768010] printk: console [ttyS0] enabled
[    0.790726] random: get_random_bytes called from 0x0000000080019706 with crng_init=0
[    0.830860] devtmpfs: mounted
[    0.836956] Freeing unused kernel memory: 84K
[    0.837125] This architecture does not have kernel memory protection.
[    0.837376] Run /sbin/init as init process

### /sbin/init ###

/sbin/init: Hello world ! (0)
/sbin/init: Hello world ! (1)
/sbin/init: Hello world ! (2)
```

QEMU + Busybox

```
[    0.000000] Linux version 5.1.0-rc2-71673-ga4fabb1ca9dd (damien@washi) (gcc version 8.2.0 (Buildroot 2018.11-rc2-00003-ga0787e9)) #64 SMP Fri Apr 19 15:06:30 JST 2019
[    0.000000] Initial ramdisk at: 0x(____ptrval____) (953856 bytes)
[    0.000000] Zone ranges:
[    0.000000]   DMA32    [mem 0x0000000080000000-0x00000000ffffffff]
[    0.000000]   Normal   empty
[    0.000000] Movable zone start for each node
[    0.000000] Early memory node ranges
[    0.000000]   node   0: [mem 0x0000000080000000-0x00000000ffffffff]
[    0.000000] Initmem setup node 0 [mem 0x0000000080000000-0x00000000ffffffff]
[    0.000000] elf_hwcap is 0x112d
...
[    0.040689] smp: Bringing up secondary CPUs ...
[    0.044944] smp: Brought up 1 node, 2 CPUs
...
[    0.147289] Unpacking initramfs...
[    0.219832] Freeing initrd memory: 928K
[    0.241662] workingset: timestamp_bits=62 max_order=12 bucket_order=0
[    0.383700] Serial: 8250/16550 driver, 1 ports, IRQ sharing disabled
[    0.397472] 10000000.uart: ttyS0 at MMIO 0x10000000 (irq = 10, base_baud = 230400) is a 16550A
[    0.457009] printk: console [ttyS0] enabled
[    0.464551] random: get_random_bytes called from 0x0000000080019706 with crng_init=0
[    0.489560] devtmpfs: mounted
[    0.491535] Freeing unused kernel memory: 84K
[    0.491704] This architecture does not have kernel memory protection.
[    0.492008] Run /sbin/init as init process
Starting rcS...
++ Mounting filesystem

# cat /proc/cpuinfo
processor       : 0
hart                            : 0
isa                             : rv64imafdcu
mmu                             : sv48

processor       : 1
hart                            : 1
isa                             : rv64imafdcu
mmu                             : sv48
```

Kendryte K210 SoC + Busybox

Sipeed MAIX Go Board (6+2 MB SRAM)

```
[    0.000000] Linux version 5.1.0-rc5-00314-g375c2321604f (damien@washi) (gcc version 8.2.0 (Buildroot 2018.11-rc2-00003-ga0787e9)) #221 SMP Fri May 10 15:17:17 JST 2019
[    0.000000] earlycon: sbi0 at I/O port 0x0 (options '')
[    0.000000] printk: bootconsole [sbi0] enabled
[    0.000000] initrd not found or empty - disabling initrd
[    0.000000] Zone ranges:
[    0.000000]   DMA32    [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000]   Normal   empty
[    0.000000] Movable zone start for each node
[    0.000000] Early memory node ranges
[    0.000000]   node   0: [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000] Initmem setup node 0 [mem 0x0000000080000000-0x00000000807fffff]
[    0.000000] elf_hwcap is 0x112d
[...]
[    0.000000] Built 1 zonelists, mobility grouping off.  Total pages: 2020
[    0.000000] Kernel command line: console=hvc0 earlycon=sbi init=/bin/bash
[    0.000000] Dentry cache hash table entries: 1024 (order: 1, 8192 bytes)
[    0.000000] Inode-cache hash table entries: 512 (order: 0, 4096 bytes)
[    0.000000] Sorting __ex_table...
[    0.000000] Memory: 6284K/8192K available (920K kernel code, 101K rwdata, 158K rodata, 393K init, 95K bss, 1908K reserved, 0K cma-reserved)
[    0.000000] SLUB: HWalign=64, Order=0-3, MinObjects=0, CPUs=2, Nodes=1
[    0.000000] rcu: Hierarchical RCU implementation.
[    0.000000] rcu: RCU calculated value of scheduler-enlistment delay is 25 jiffies.
[    0.000000] NR_IRQS: 0, nr_irqs: 0, preallocated irqs: 0
[...]
[    0.251433] Freeing unused kernel memory: 392K
[    0.254361] This architecture does not have kernel memory protection.
[    0.259473] Run /bin/bash as init process

BusyBox v1.30.1 (2019-05-10 14:49:46 JST) hush - the humble shell

# mount -t proc none /proc
# cat /proc/cpuinfo
processor : 0
hart : 0
isa : rv64imafdc

processor : 1
hart : 1
isa : rv64imafdc
```

Western Digital®