Consolidating representations of the physical memory

Mike Rapoport
<rppt@linux.ibm.com>
Memory detection

**Firmware**

- Detect memory
- Reserve for own use
  - E.g. ACPI tables, EFI services
- Create memory map
  - HWRPB, ATAGs, DT, e820, ...
- Create NUMA info
  - DT, SRAT, PCI config space, ...
Memory detection

Early boot

- Parse memory map
  - Reserve memory used by firmware
  - Create arch-specific representation
  - Pass memory ranges to memblock

- Setup NUMA topology
  - Verify NUMA tables supplied by firmware
  - Assign memory to nodes
  - Create arch-specific representation
Early boot

- Memory resources initialization
  - Initialize “System RAM” resource
  - Add memory resources for reserved regions
  - Add memory resources for kernel image

- Setup of memory sysfs

Memory detection

- System RAM
- System RAM
- System RAM
- System RAM
- ACPI Tables
- ACPI NVS
- Reserved
- Unknown E820 type
- System RAM
- Reserved
- System RAM
- Reserved
- System RAM
- Kernel code
- Kernel rodata
- Kernel data
- Kernel bss
Run time memory range checks and operations

- Arch-specific representation
  - x86: e820__mapped_any(), phys_to_target_node()
  - ppc: for_each_drmem_lmb()
  - s390: oldmem

- Resource tree
  - page_is_ram(), walk_system_ram_range()

- Memblock
  - memblock_is_map_memory()
kexec

- Relies on `/proc/iomem`
  - Find memory to load kernel image
  - Create memory map for the second kernel
- Only x86 can use `/sys/firmware/memory`
  - Except on Xen and some EFI systems
Create a generic representation of the physical memory

- Consistent with firmware memory map
- Backward compatible with existing uABI
- Flexibility to accommodate architecture-specific needs
Alternatives

- New data structure
- Resource tree
- memblock
Proposal

Use memblock as physical memory representation

- Enable ARCH_KEEP_MEMBLOCK for NUMA architectures
- Generate “System RAM” from memblock
  - [https://lore.kernel.org/all/20210531122959.23499-1-rppt@kernel.org/](https://lore.kernel.org/all/20210531122959.23499-1-rppt@kernel.org/)
- Close the gap in e820 vs memblock integration
- Rename `memblock_alloc()` back to `bootmem_alloc()`?
Thank you!