FUJITSU
shaping tomorrow with you

Memory Hotplug

May 29th, 2013

Yasuaki Ishimatsu <isimatu.yasuaki@jp.fujitsu.com>
FUJITSU LIMITED

Agenda

FUJITSU

- Motivation of memory hotplug
- What is memory hotplug?
- Development of memory hotplug
- To-do Lists

FUJITSU

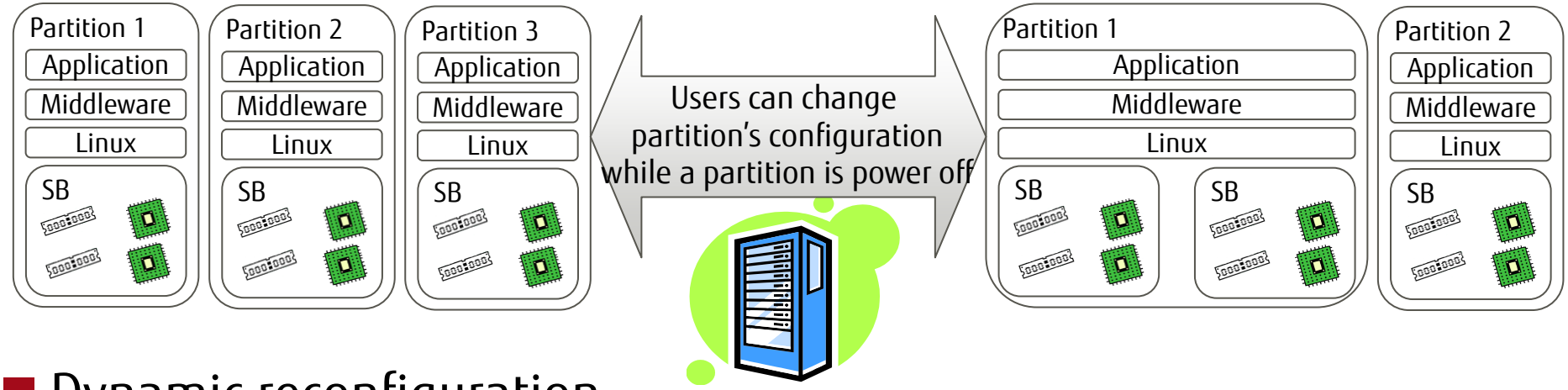
MOTIVATION OF MEMORY HOTPLUG

Hardware partitioning & Dynamic reconfiguration

FUJITSU

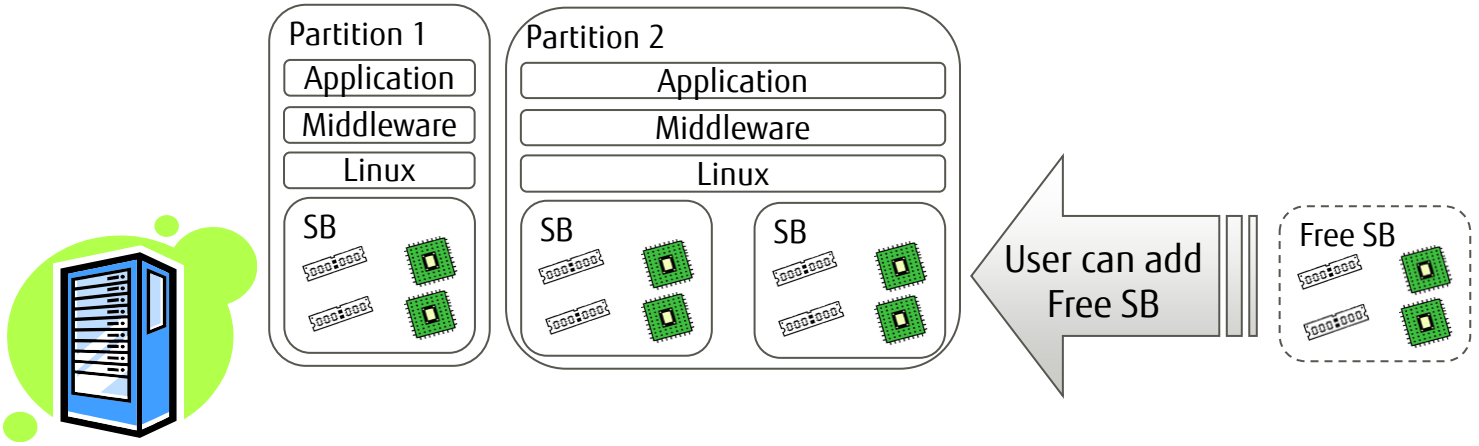
Hardware partitioning

- System can have several servers in a box with flexibility



Dynamic reconfiguration

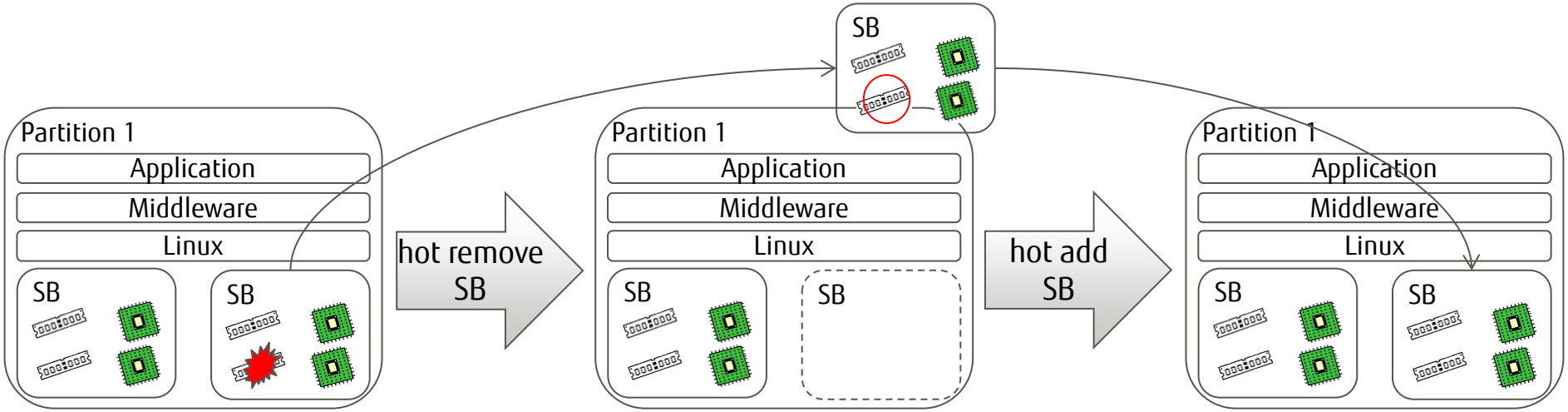
- System can change partition's configuration at runtime



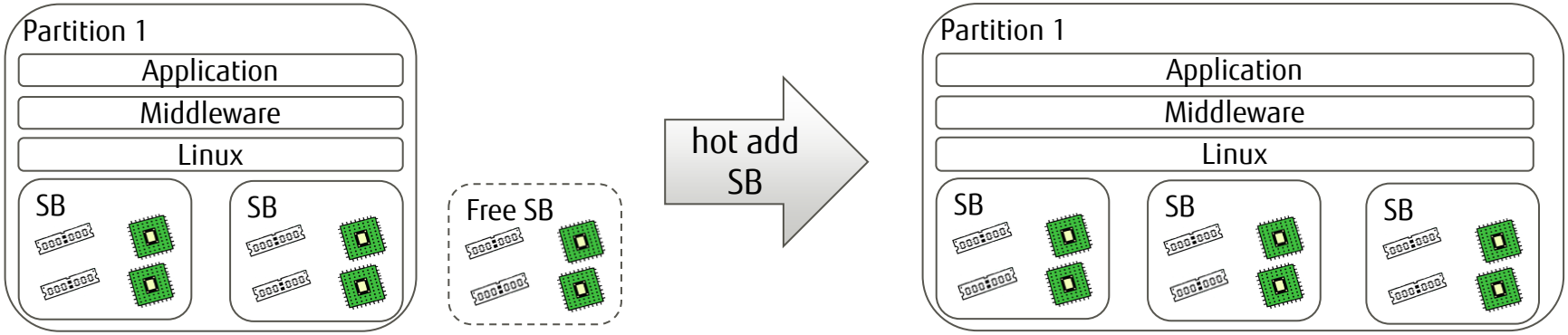
Purpose of memory hotplug (1)

FUJITSU

- Dynamic reconfiguration enhance RAS



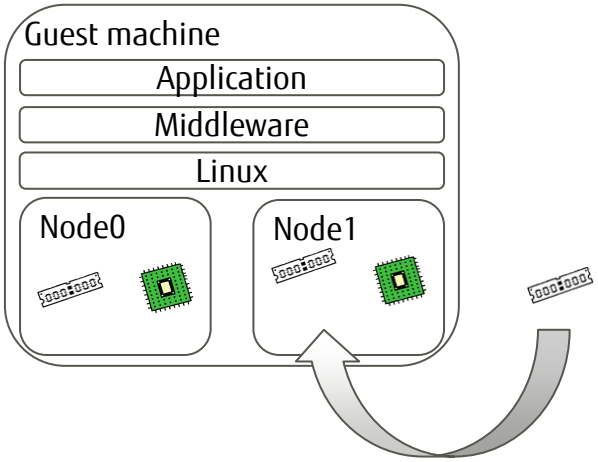
- Dynamic reconfiguration is used for resize



Purpose of memory hotplug (2)

FUJITSU

- Memory hotplug will be supported by KVM
  - "ACPI memory hotplug" by Vasilis Liaskovitis
    http://lists.gnu.org/archive/html/qemu-devel/2012-12/msg02693.html



⇨ Memory hotplug has been supported by several OSs. But Linux has not copmletely support it yet.

FUJITSU

WHAT IS MEMORY HOTPLUG

Memory hotplug

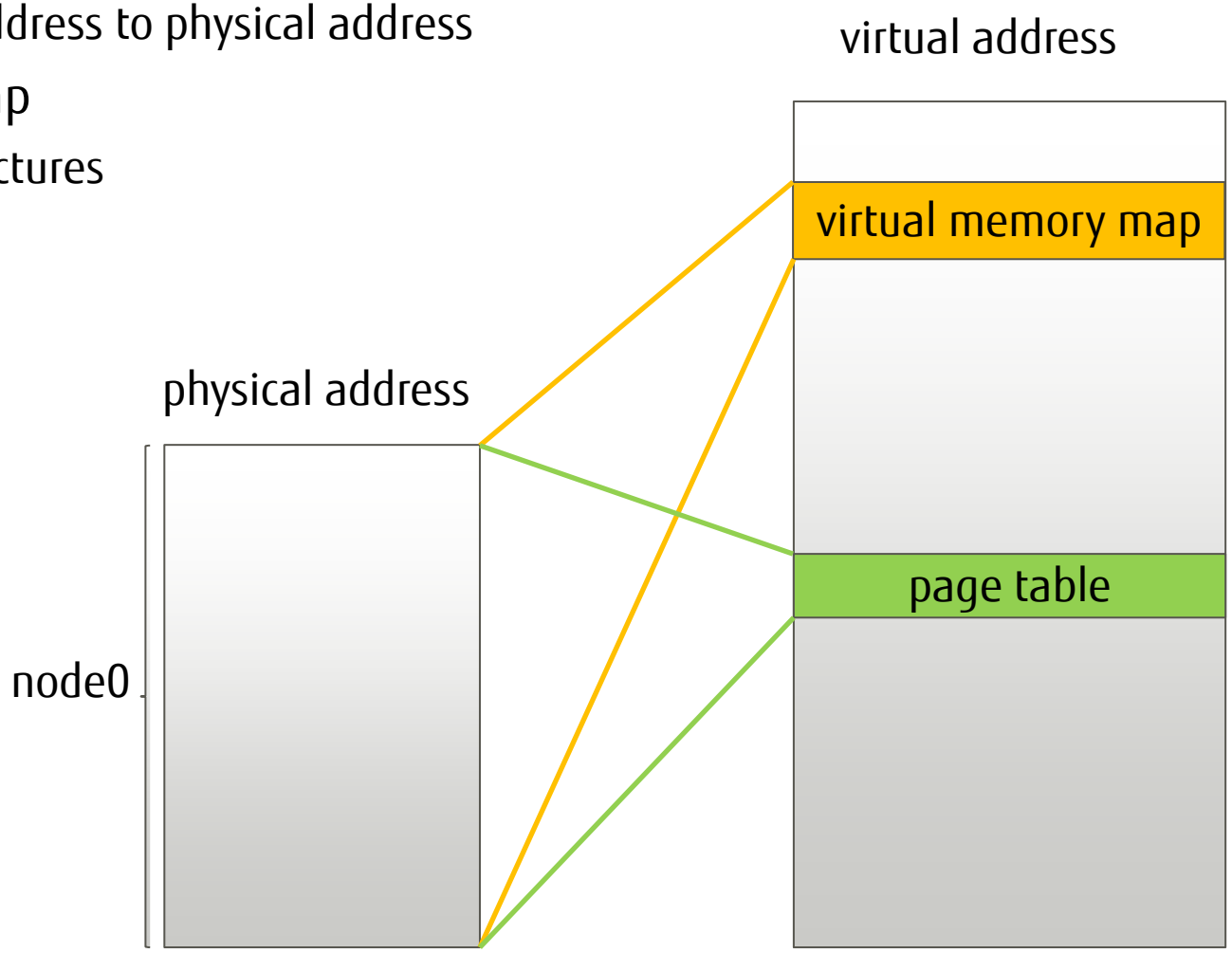
FUJITSU

- Memory hotplug allows to users to increase/decrease the amount of memory
  - Physical memory hotplug phase (For hot adding/removing DIMMs physically)
    - memory hot add
    - memory hot remove
  - Logical memory hotplug phase (For changing the amount of memory)
    - memory online
    - memory offline

Memory management

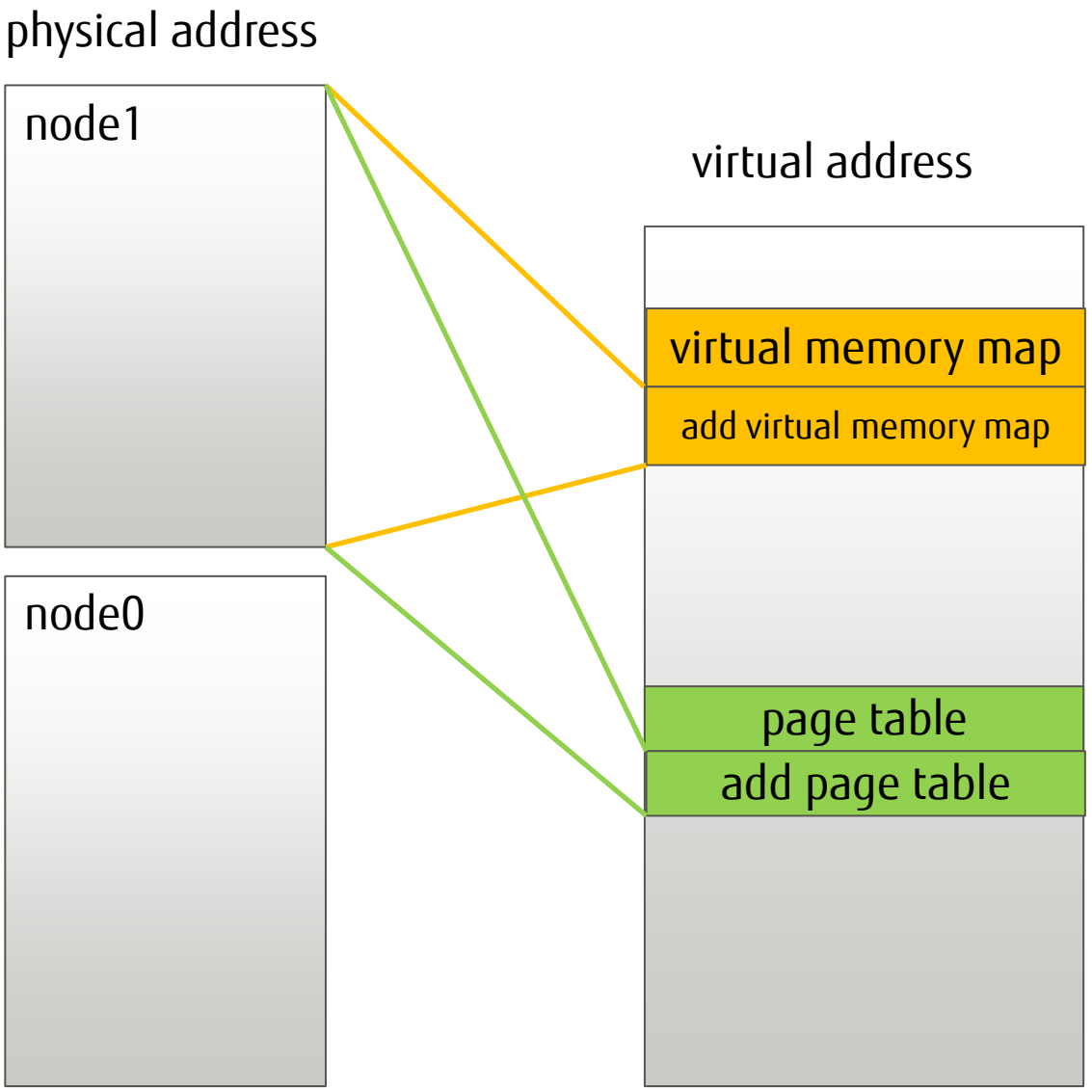
FUJITSU

- Kernel manages physical memory by using
  - page table (direct mapping)
    - calculate virtual address to physical address
  - virtual memory map
    - manage page structures



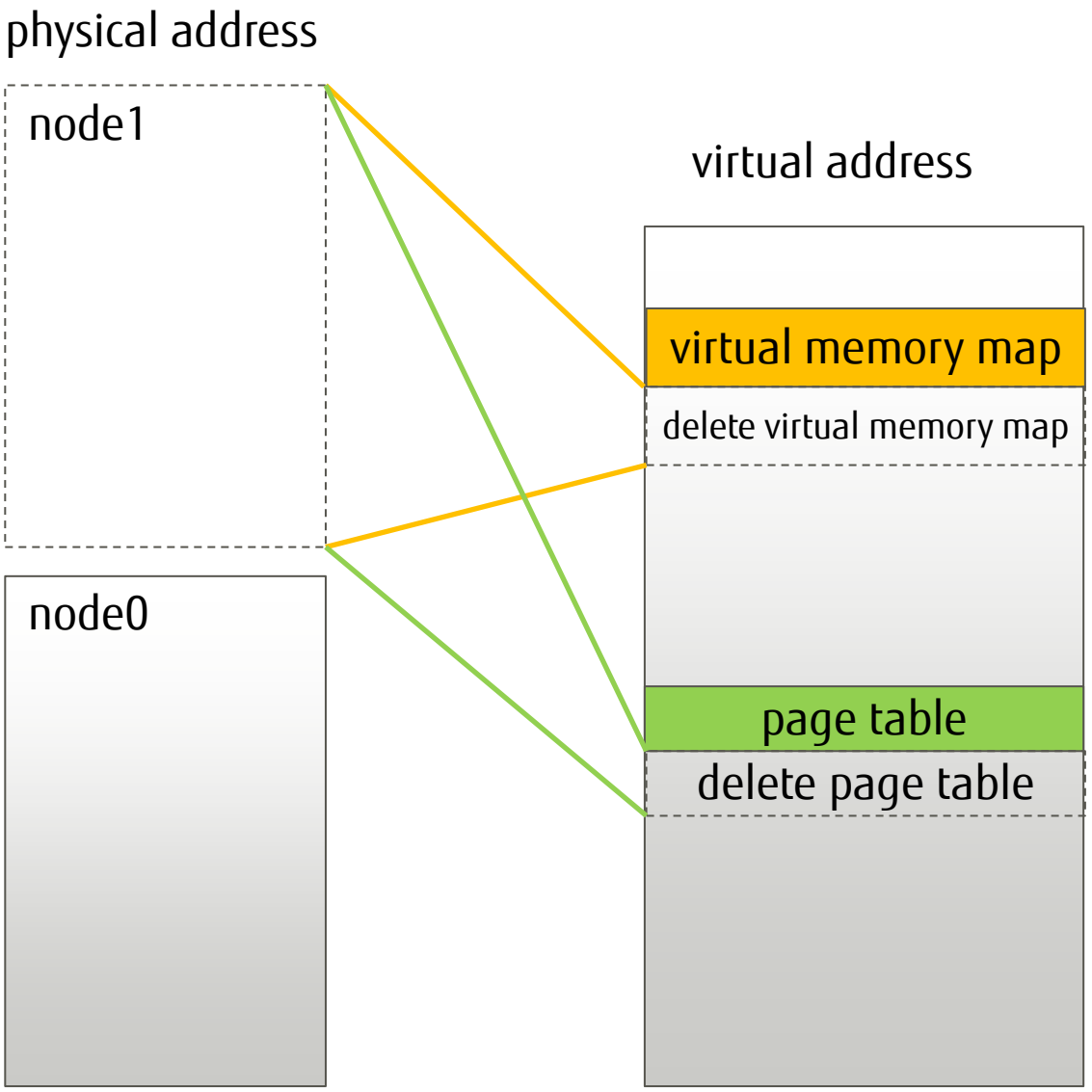
Memory hot add

- Prepare following items for managing added memory
  - page table (direct mapping)
  - virtual memory map



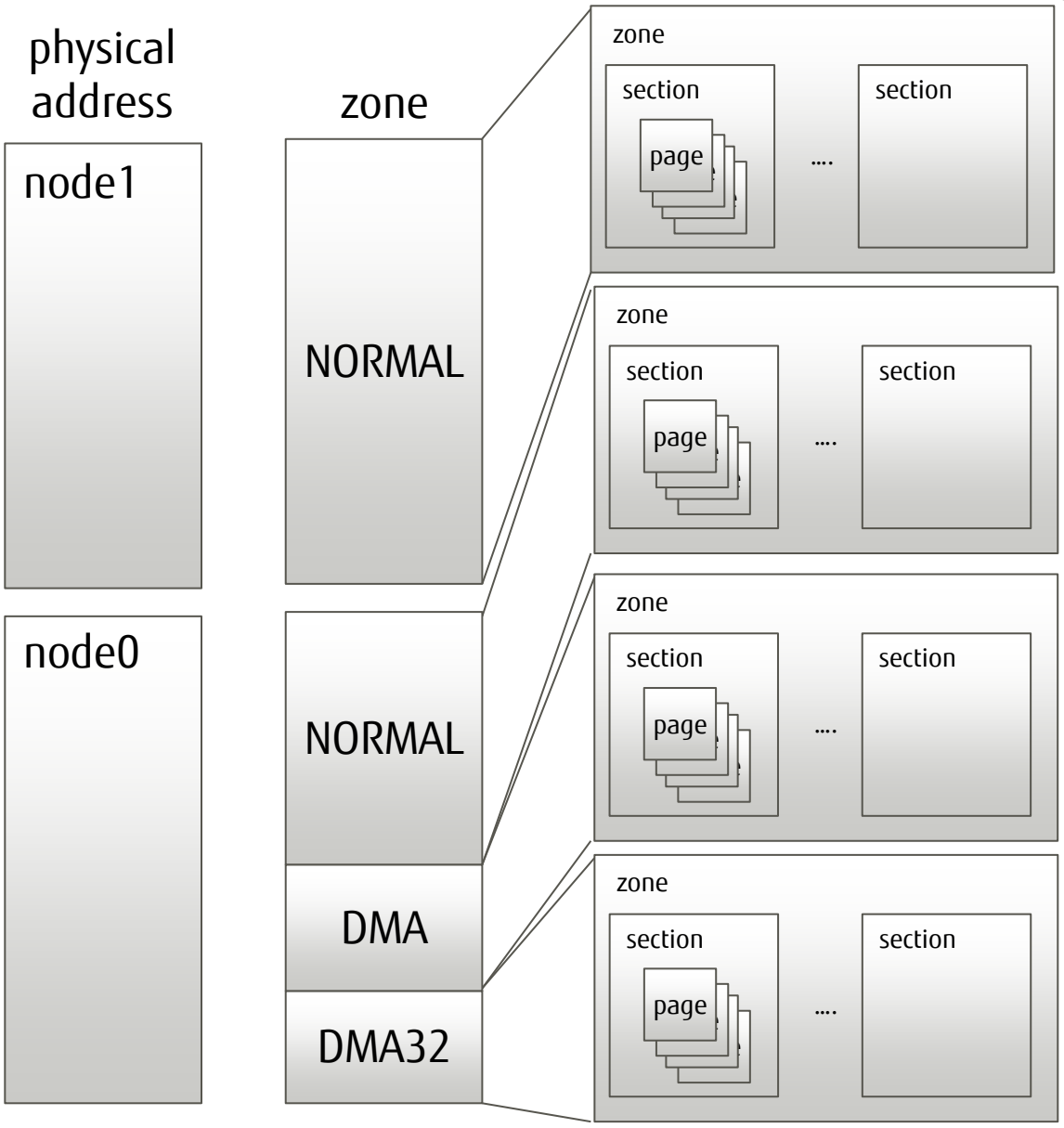
Memory hot remove

- Delete following items for managing removed memory
  - page table (direct mapping)
  - virtual memory map



Online memory & Offline memory

- All pages are managed by each zone structures
- User can online/offline memory by echoing sysfs file
  - echo online/offline > /sys/devices/system/memory/memoryX/state
- Online memory
  - change page state to usable
- Offline memory
  - change page stat to unusable



FUJITSU

DEVELOPMENT OF MEMORY HOTPLUG

Status of memory hotplug on linux-3.2

FUJITSU

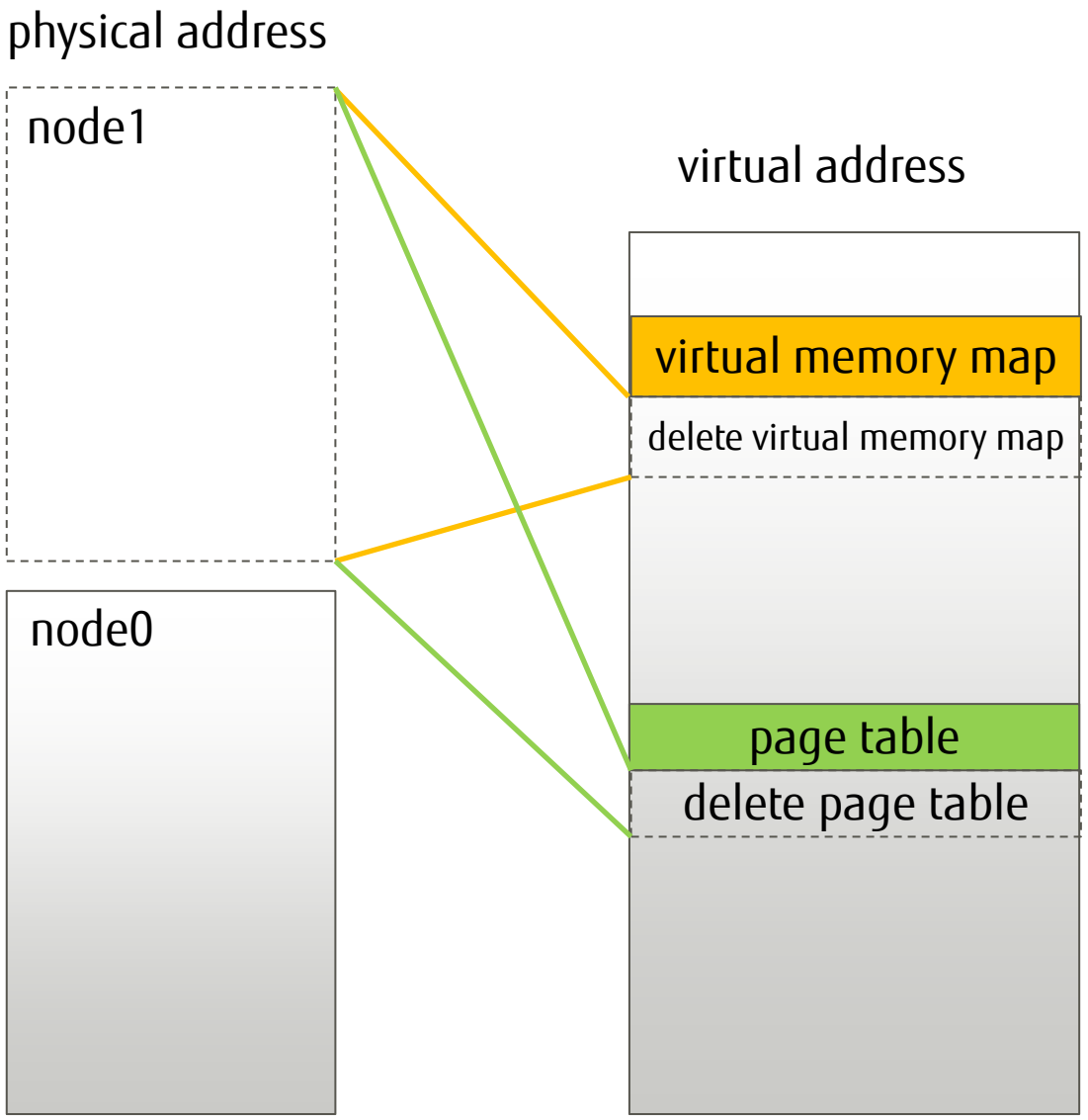
- Hot add memory
  - Support
- Hot remove memory
  - Not supoprt
- Online memory
  - Support
- Offline memory
  - Support with limitation

Required items for memory hot remove

FUJITSU

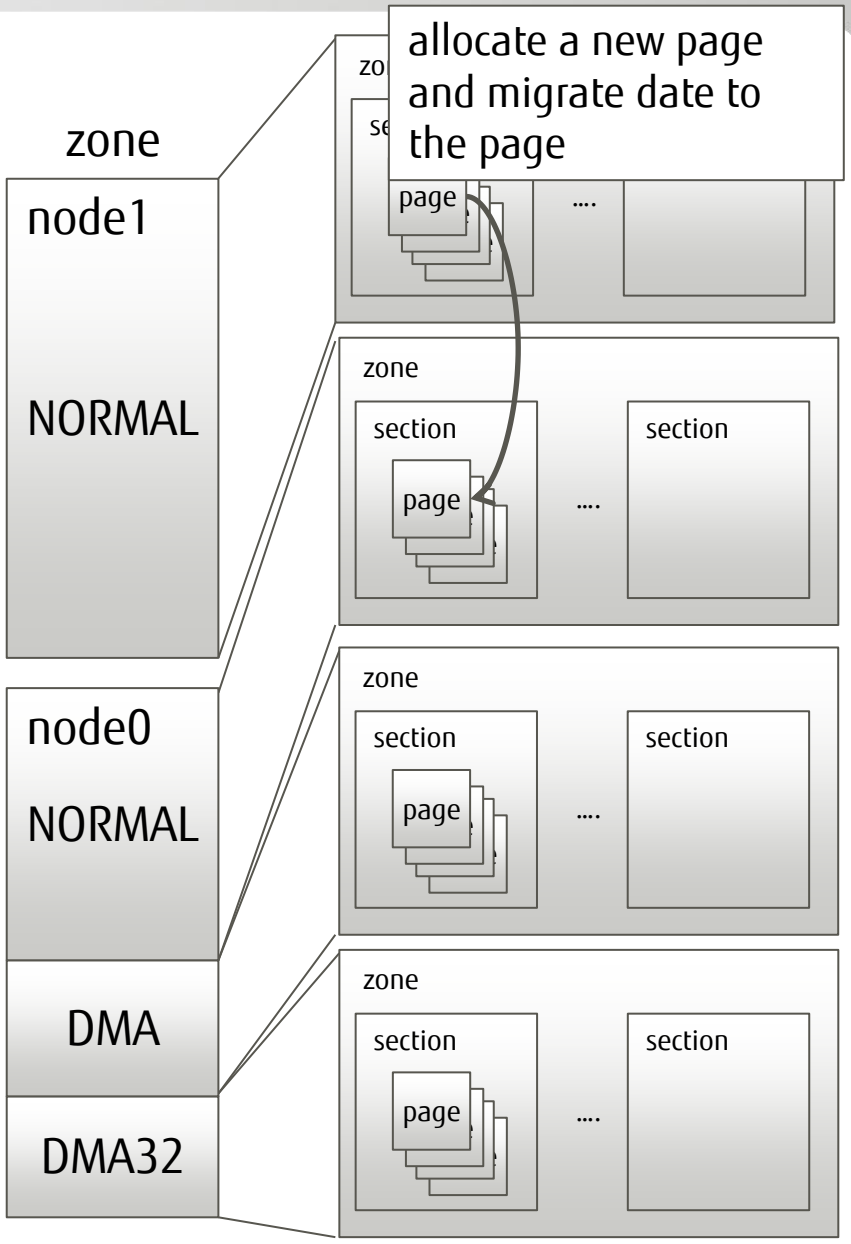
- Delete following items for managing removed memory
  - page table (direct mapping)
  - virtual memory map
  - /sys/firmware/memmap sysfs
  - /sys/devices/syste/node/memoryX sysfs
- Improve ACPI memory hotplug frame work
- Update zone information

All of them are merged into linux3.9



Limitation of memory offline

- When offlining memory has data, the data is migrated to other page
- But migratable pages are only page cache and anonymous page, called movable memory
- So if memory is used as other purpose except for movable memory, called kernel memory, the memory cannot be offlined



Comparison of 2 ways of offlining

FUJITSU

- Support to offline kernel memory
  - Pros.
    - All memory can be offlined
    - No performance impact caused by NUMA
  - Cons.
    - Kernel address (P<->V relationship) should change completely
- Make a node only of movable memory
  - Pros.
    - We can make use of existing feature
  - Cons.
    - The node can use as only page cache and anonymous page
    - Performance impact caused by NUMA

Comparison of 2 ways of offlining

FUJITSU

- Support to offline kernel memory
  - Pros.
    - All memory can be offlined
    - No performance impact caused by NUMA
  - Cons.
    - Kernel address (P<->V relation ship) should change completely

For supporting kernel memory offline, it will take several years.

Comparison of 2 ways of offlining

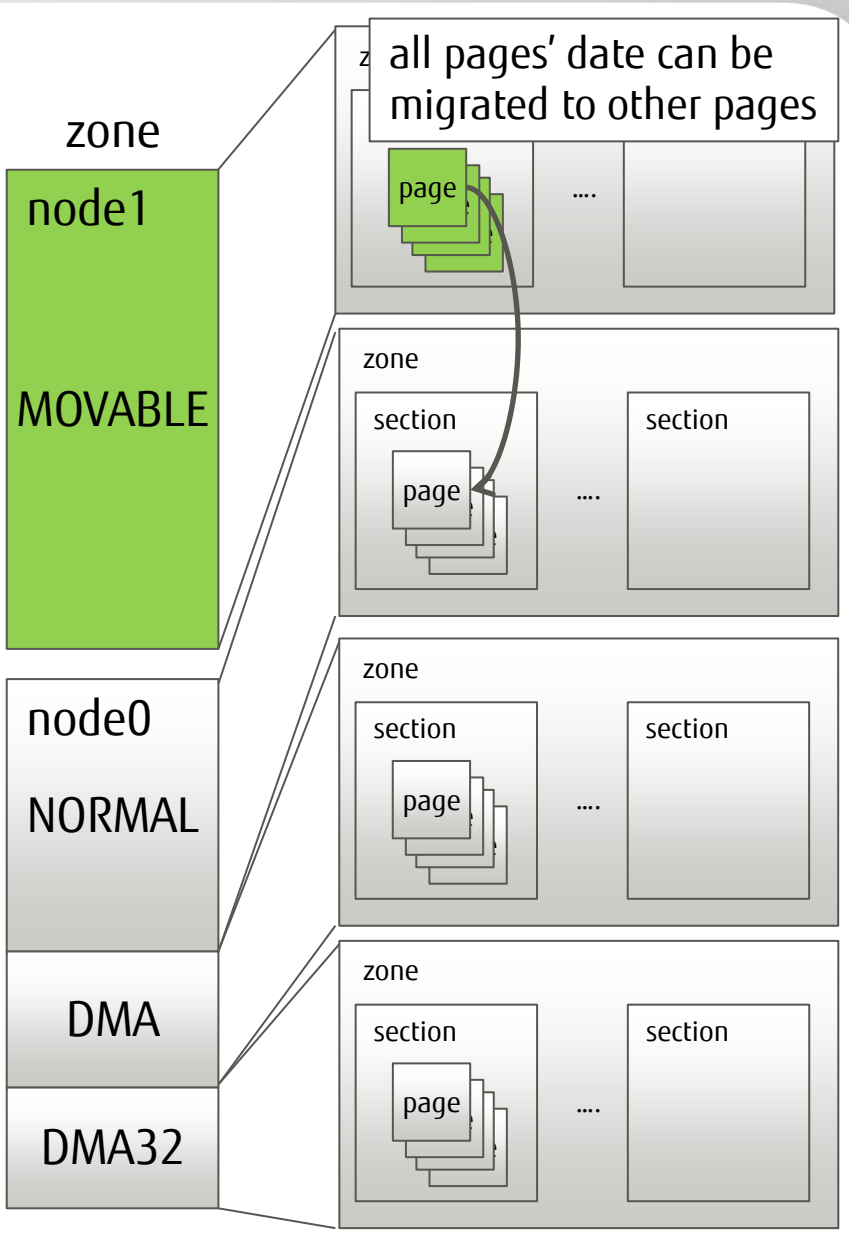
FUJITSU

As first step, we develop following solution

- Make a node consists only of movable memory (movable node)
  - Pros.
    - We can make use of existing feature
  - Cons.
    - The node can use as only page cache and anonymous page
    - Performance impact by NUMA

Make a node consists only of movable memory

- For making a memory range of movable memory, Linux has ZONE_MOVABLE, this is not created automatically
- If a node is constructed by only ZONE_MOVABLE, the whole memory on the node can be migrated to other memory and offlined



ZONE_MOVABLE configuration

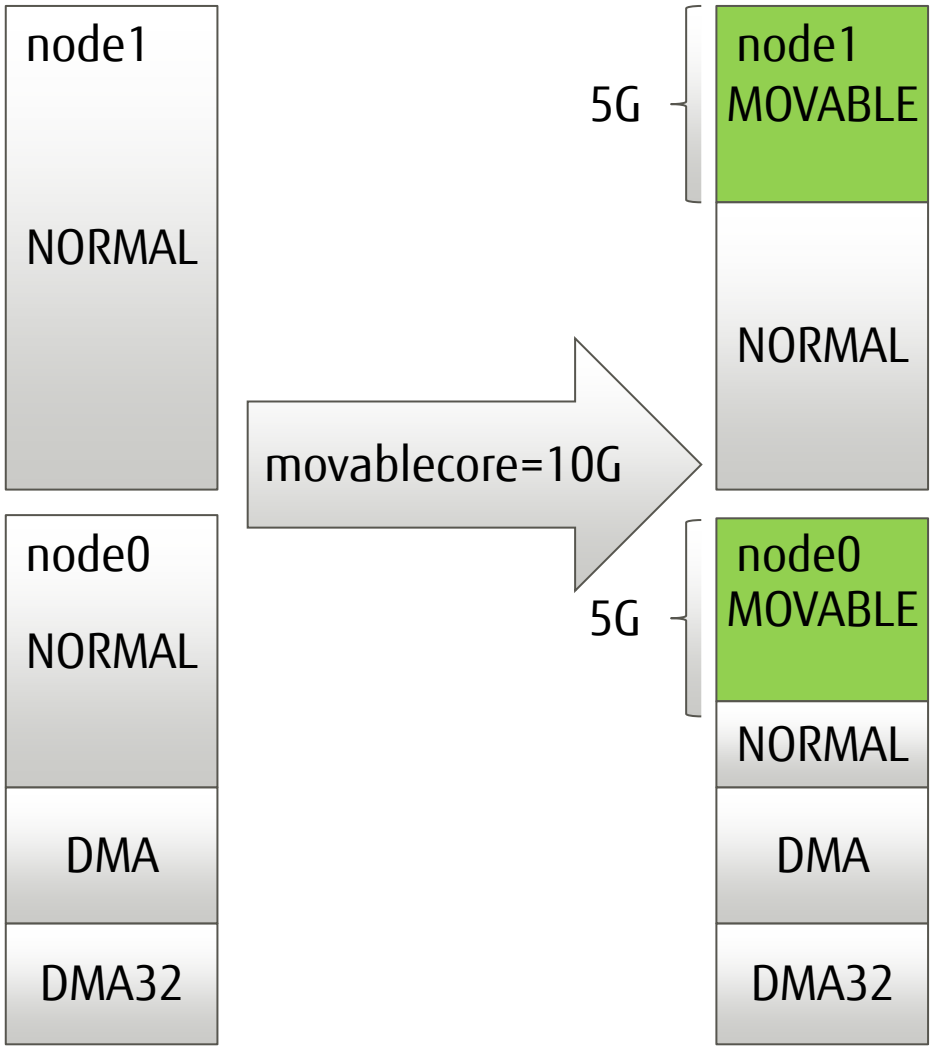
FUJITSU

- Enhance/Add features for creating ZONE_MOVABLE
  1. New boot option, movablecore=acpi
  2. New “online” feature, online_movable

Original boot option of creating MOVABLE zone

FUJITSU

- Linux has a boot option, called movablecore=, for creating MOVABLE zone
- The boot option can specifies amount of movable memory in a system
- But movable memory is evenly distributed to all nodes.

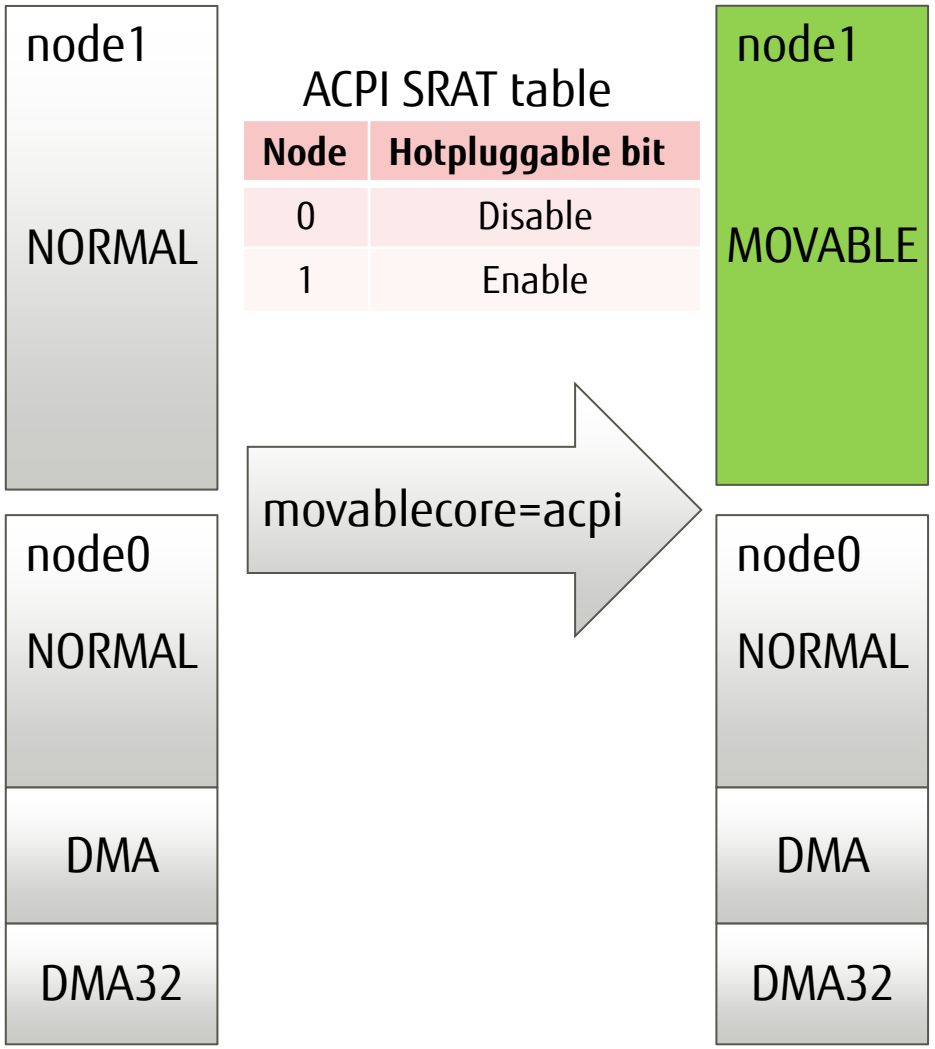


New boot option

FUJITSU

- We are proposing a new boot option “movablecore=acpi”
  - use memory affinity structure in SRAT table
  - if hotpluggable bit is enable, the memory range is managed to MOVABLE zone

The feature is under developing

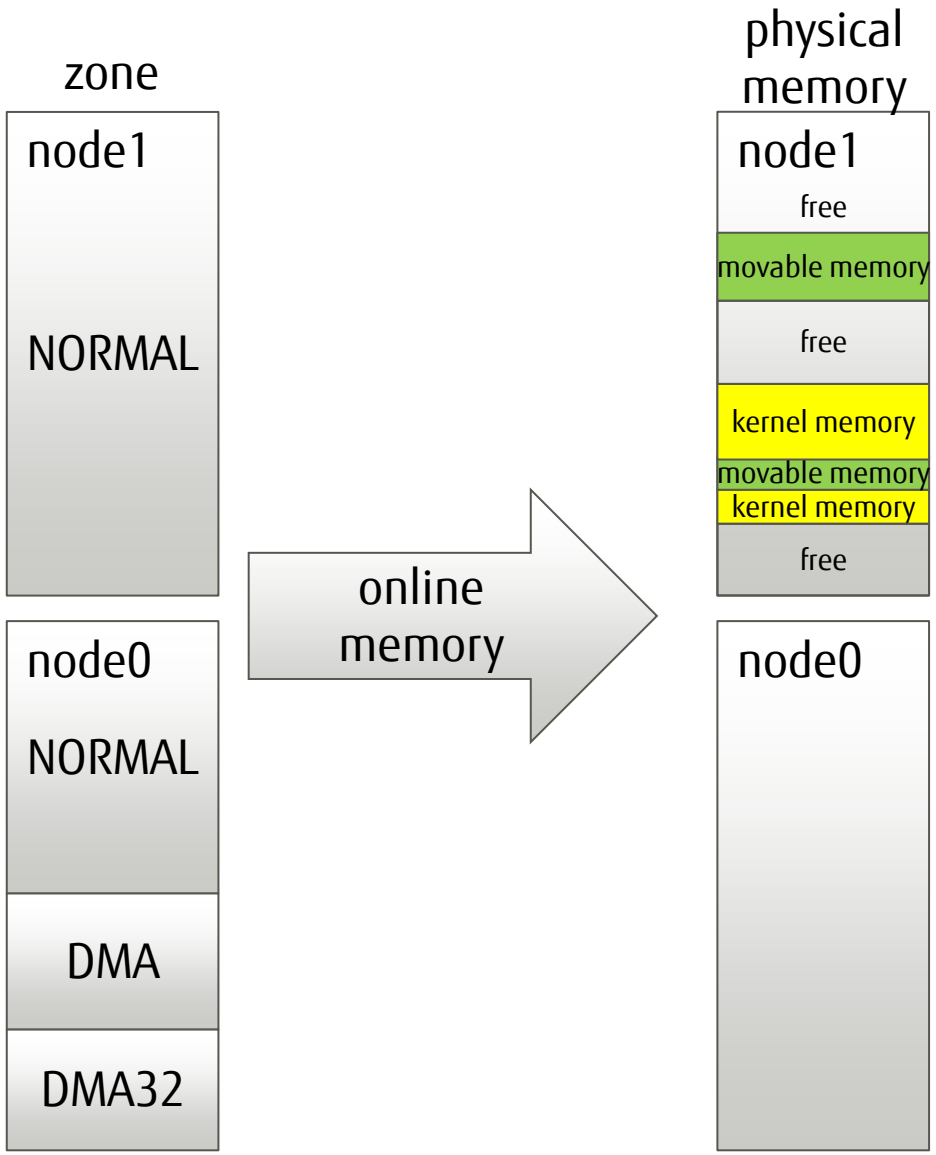


Lack of feature at onlining memory

FUJITSU

- Hot added memory is always managed by NORMAL zone
- When onlining memory, the memory may contains kernel memory and movable memory

  echo online > /sys/devices/system/node/nodeX/memoryY/state



New interface at memory onlining

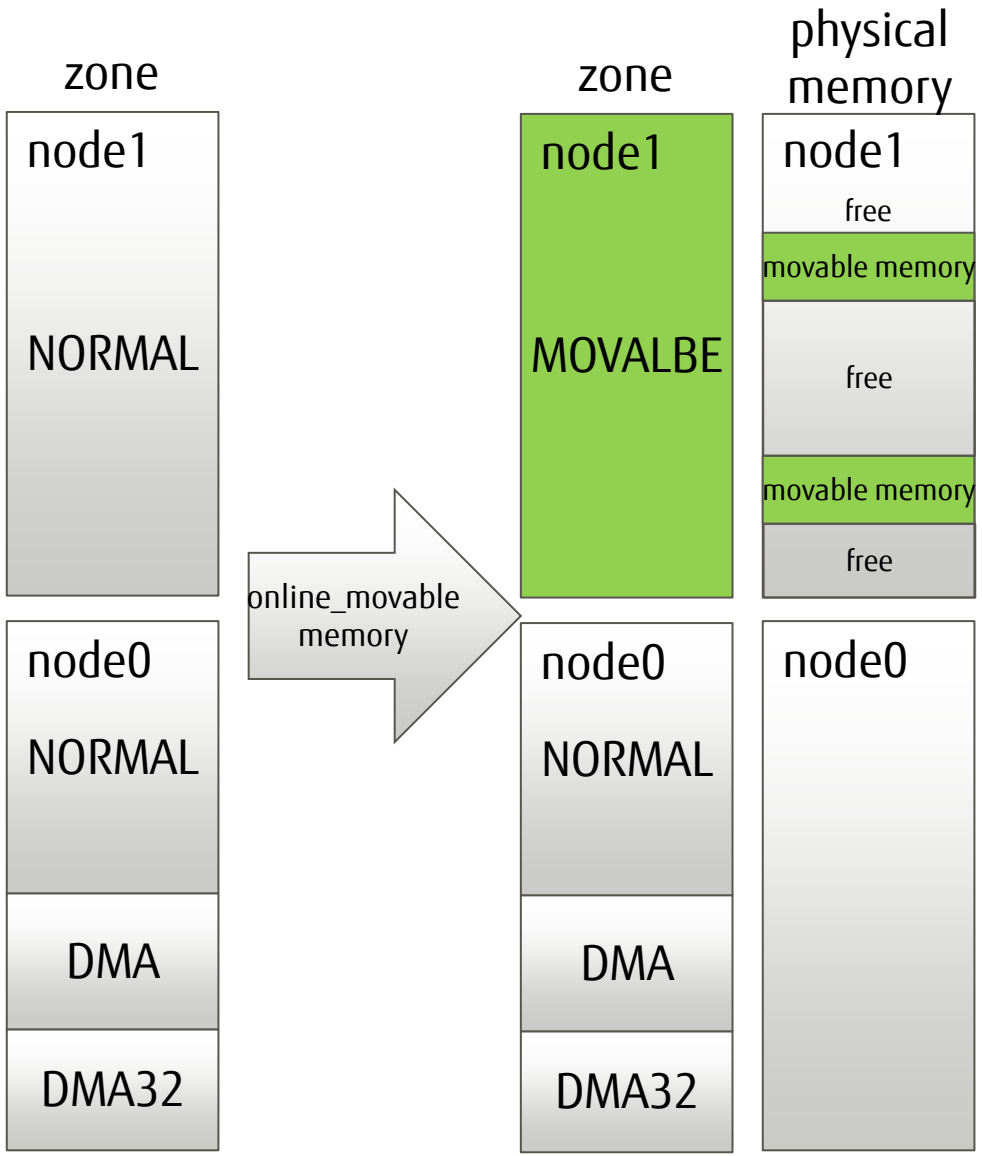
FUJITSU

- Online memory into ZONE_MOVABLE

  echo online_movable > /sys/devices/system/node/nodeX/memoryY/state

  ⇩

  The feature is merged into linux3.8



Current status of memory hotplug

FUJITSU

- Hot add memory
  - Support
- Hot remove memory
  - supoprt
- Online memory
  - Support
- Offline memory
  - Support with limitation
    - added online_movable option
    - under developing movablecore=acpi option

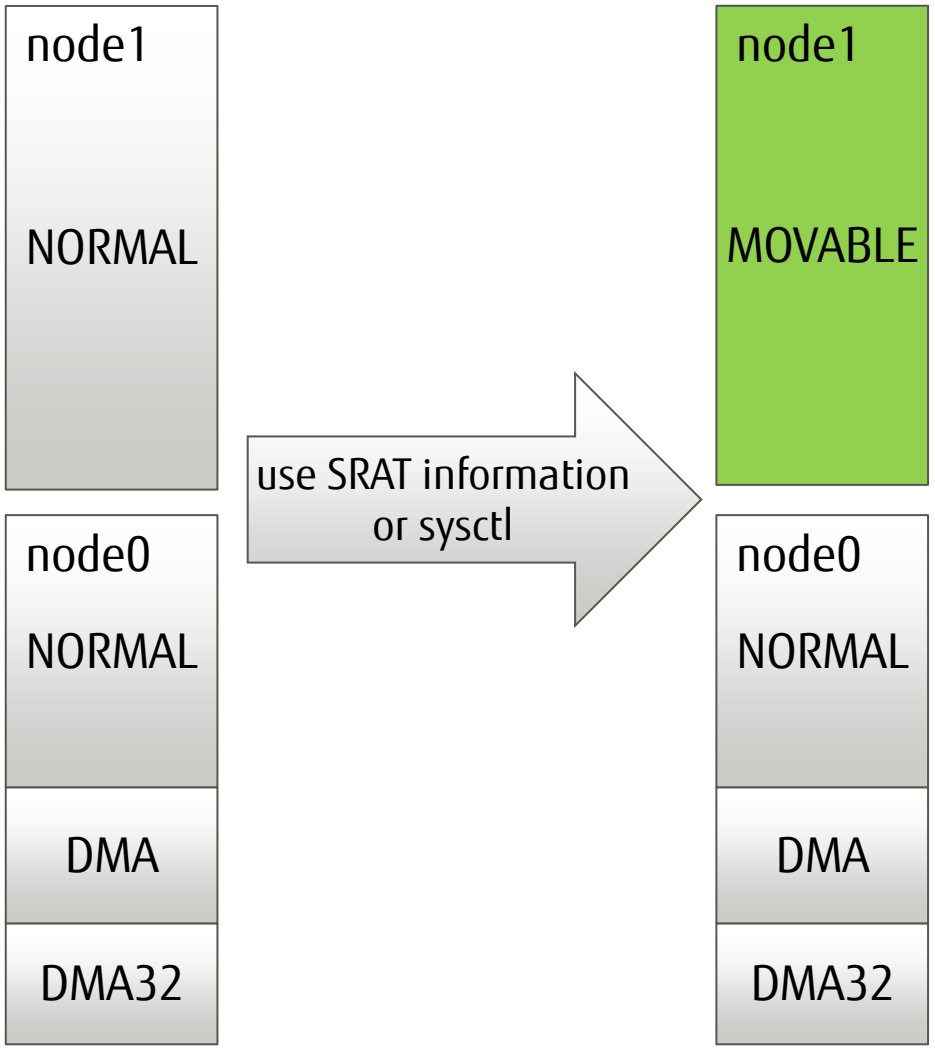
FUJITSU

TO-DO LISTS

Switch of changing hot added memory's zone

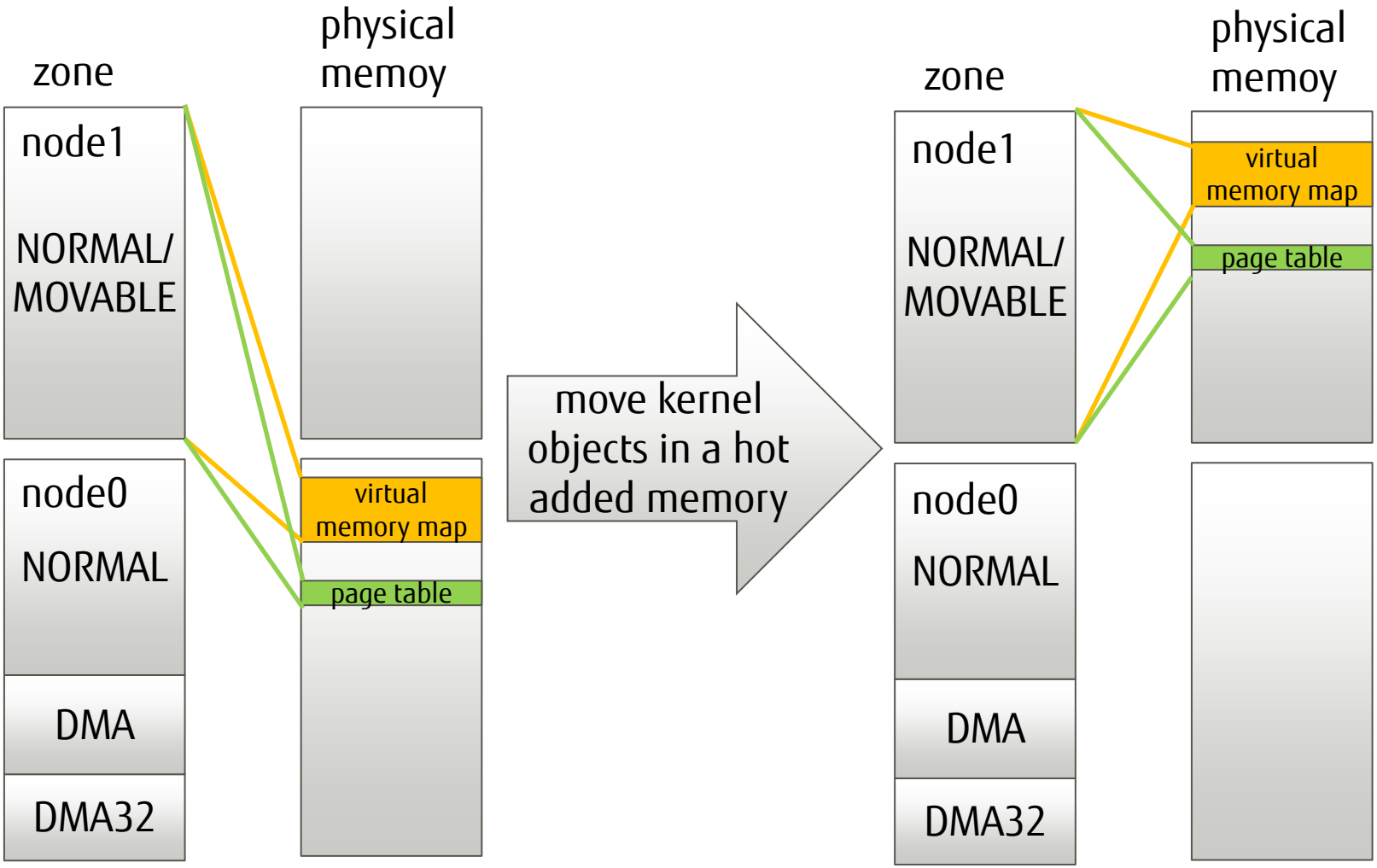
FUJITSU

- Hot added memory is always managed by NORMAL zone
- If user want to hot remove memory, user need to use:
  - echo online_movable > /sys/devices/system/nodeX/memoryY/state
- Prepare
  - SRAT information
    - If movablecore=acpi is defined, check hotpluggable bit of SRAT information
  - sysctl
    - vm.hotadd_memory_treat_as_movable



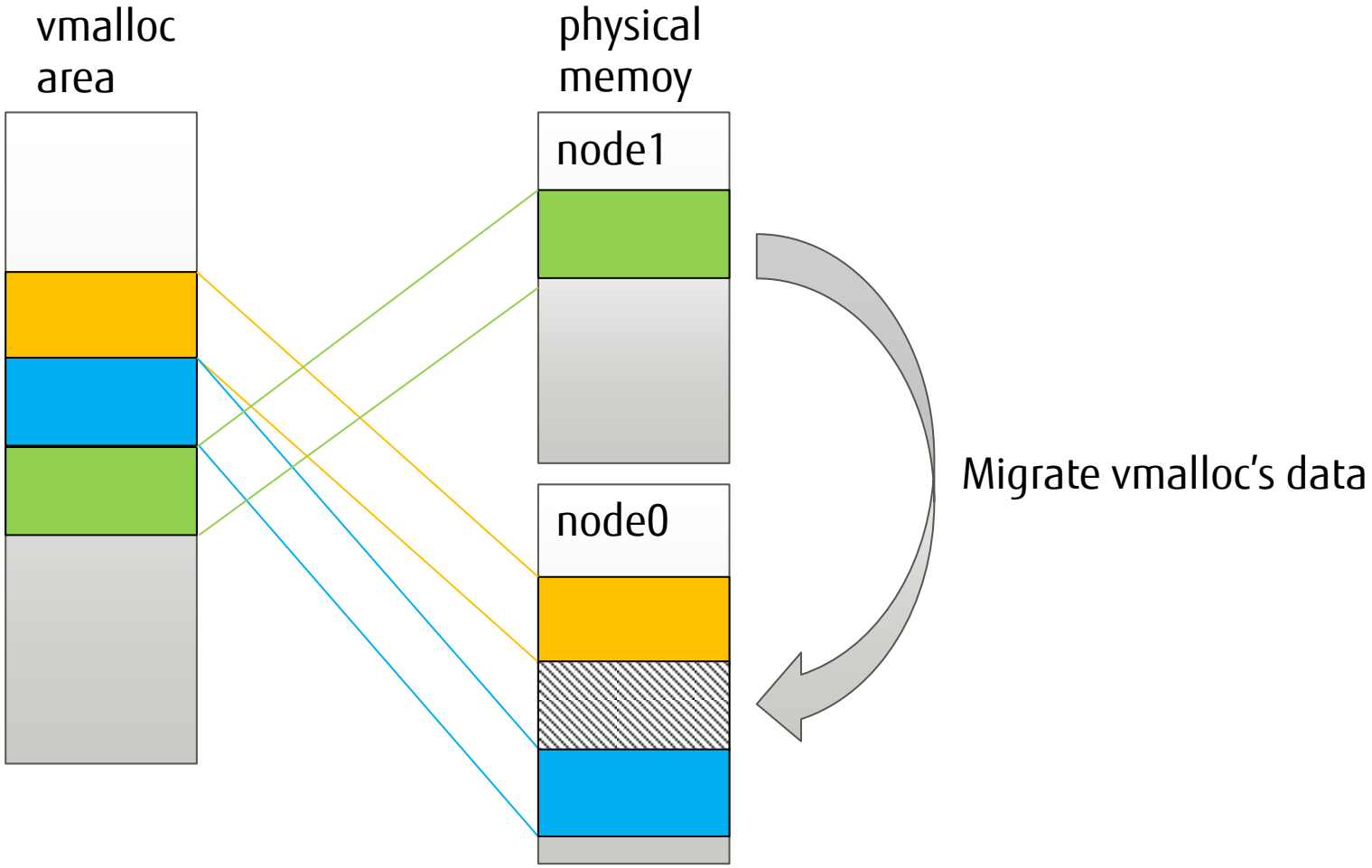
Kernel object in a hot added memory

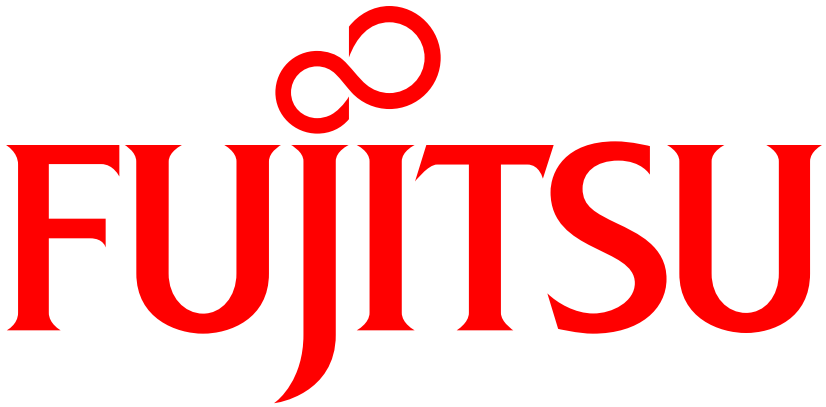
- When hot adding memory, kernel objects like page table and virtual memory map are allocated into other memory.



Migrate vmalloc area

- Vmalloc area is not continuous physically
- When offlining vmalloc's region, the data on the range is migrated to other memory





shaping tomorrow with you