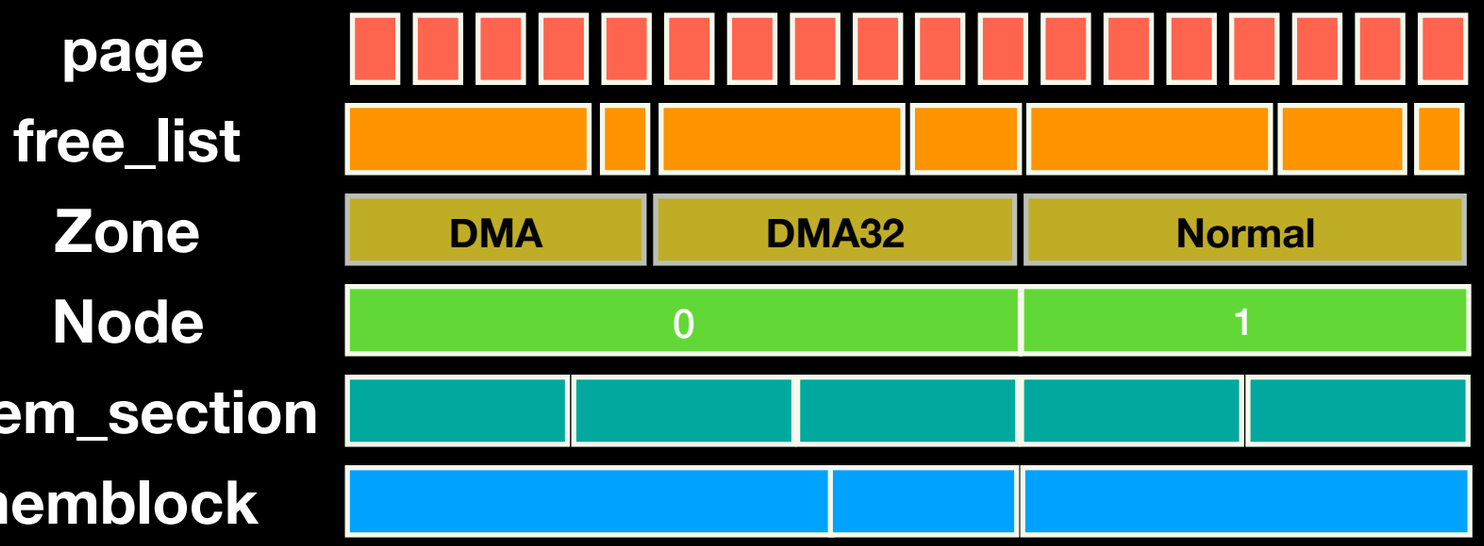
内存管理的层次结构

Wei Yang
<richard.weiyang@gmail.com>

议程

- 现有的层次结构
- 分层的原因

现有的层次结构


slub
page
free_list
Zone
DMA
DMA32
Normal
Node
0
1
mem_section
memblock

内核是软件和硬件的结合

是两者妥协的产物

分层的原因

- 描述硬件属性：Node, Zone, Page
- 特性支持：hotplug
- 提升扩展性：zone->free_list，slub，pcp
- 节省空间：mem_section

slub

page

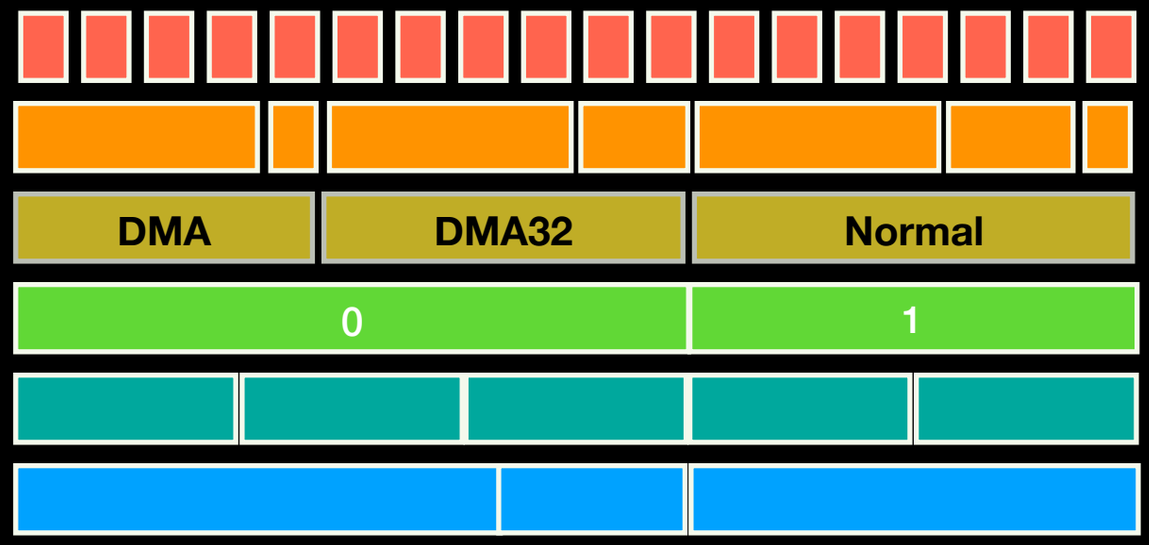
free_list

Zone

Node

mem_section

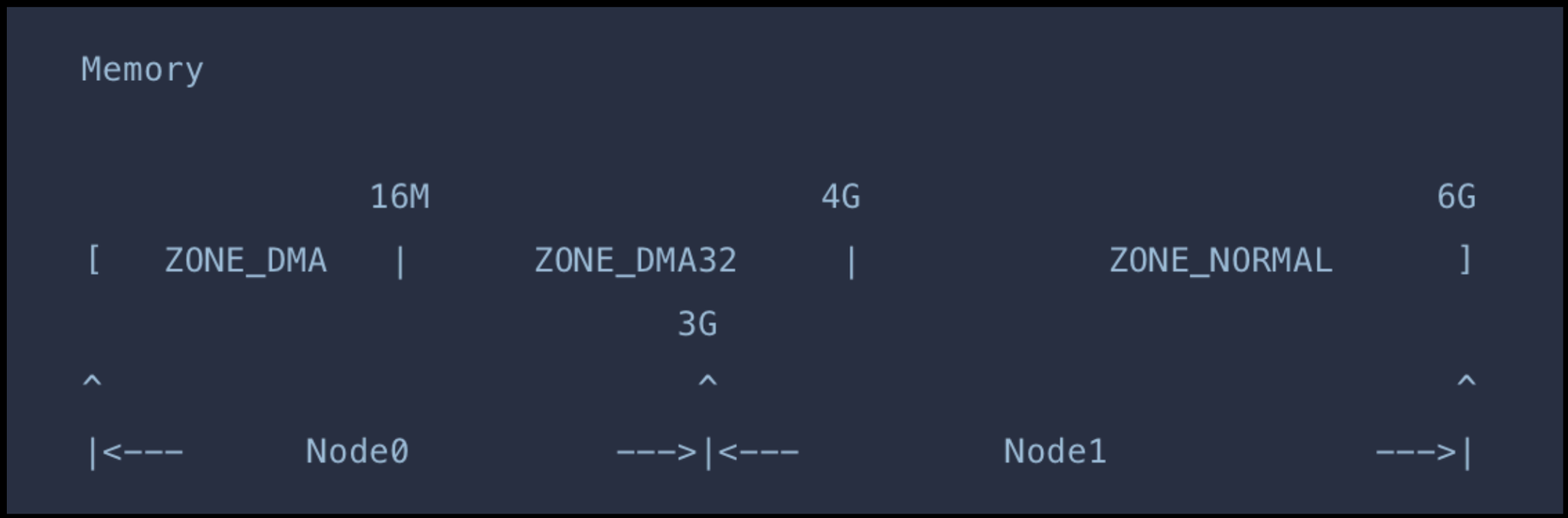
memblock



软件诉求

硬件诉求

Node 和 Zone的含义



热插拔支持

扩展性考量 1


node_data[0]
node_id
(int)
node_zones[MAX_NR_ZONES]
(struct zone)
zone_pgdat
[ZONE_DMA]
0
16M
[ZONE_DMA32]
16M
3G
[ZONE_NORMAL]
empty

扩展性考量 2


struct zone
pageset
(struct per_cpu_pageset *)
cpu0
cpu1
cpuN
...
pcp
(struct per_cpu_pages)
count
high
batch
lists[MIGRATE_PCPTYPES]

节省空间 page的存放

```
mem_section[NR_SECTION_ROOTS][SECTIONS_PER_ROOT]

= [DIV_ROUND_UP(NR_MEM_SECTIONS, SECTIONS_PER_ROOT)] [SECTIONS_PER_ROOT]

    [0]          [1]                                [SECTIONS_PER_ROOT - 1]
    +------------+------------+        +------------+------------+
[0] |            |            |   ...  |            |            |
    +------------+------------+        +------------+------------+

    +------------+------------+        +------------+------------+
[1] |            |            |   ...  |            |            |
    +------------+------------+        +------------+------------+

    +------------+------------+        +------------+------------+
[2] |            |            |   ...  |            |            |
    +------------+------------+        +------------+------------+
```