VM Overcommit

Константин Белоусов
kib@freebsd.org

26 сентября 2009 г.

Revision : 1.4

Что такое swap overcommit

malloc -> SIGSEGV ?

```
char *p;
...
p = malloc(1024);
if (p != NULL)
        p[0] = 'a';
```

Malloc implementation

malloc is implemented over sbrk(2) or mmap(2)

Malloc implementation

malloc is implemented over sbrk(2) or mmap(2)

mmap -> SIGSEGV ?

```
char *p;
...
p = mmap(0, PAGE_SIZE,
    PROT_READ|PROT_WRITE,
    MAP_PRIVATE|MAP_ANON, -1, 0);
if (p != (char *)MAP_FAILED)
        p[0] = 'a';
```

Other OSes

Solaris

Overcommit is always disabled

Linux

sysctl vm.overcommit_memory

- 0 – pretend to be reasonable
- 1 – always overcommit
- 2 – disable to overcommit more percentage of system memory then vm.overcommit_ratio

Основные структуры VM

vmspace, Process Address Space

Описывает однородные регионы адресного пространства.
Состоит из pmap и vm_map.

- pmap – данные для аппаратуры трансляции адресов
- vm_map – список регионов

Основные структуры VM

vmspace, Process Address Space

Описывает однородные регионы адресного пространства.
Состоит из pmap и vm_map.

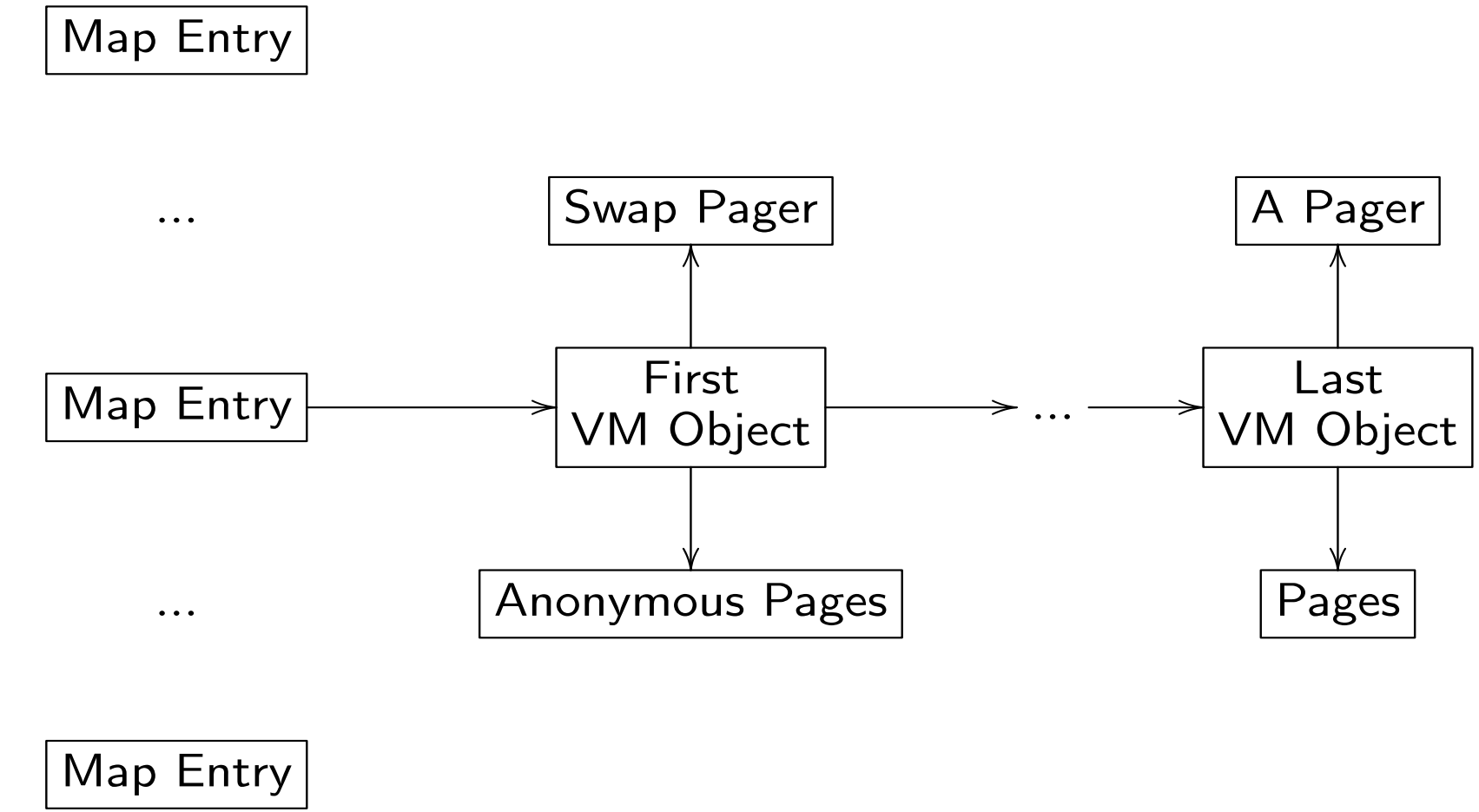
- pmap – данные для аппаратуры трансляции адресов
- vm_map – список регионов

VM Object & Pager

Контейнер для физических страниц памяти

- SWAP
- VNODE
- DEVICE
- PHYS
- SG
- ...

VM – Все вместе



Laziness

VM is eager

Чтение (с диска) выполняется заранее и большими порциями, чтобы избежать ожидания данных с диска в тот момент, когда они нужны.

Laziness

VM is eager

Чтение (с диска) выполняется заранее и большими порциями, чтобы избежать ожидания данных с диска в тот момент, когда они нужны.

VM is lazy

VM старается отложить запись на диск и вычисления на как можно более поздний срок.
Возможно, они никогда не понадобятся.

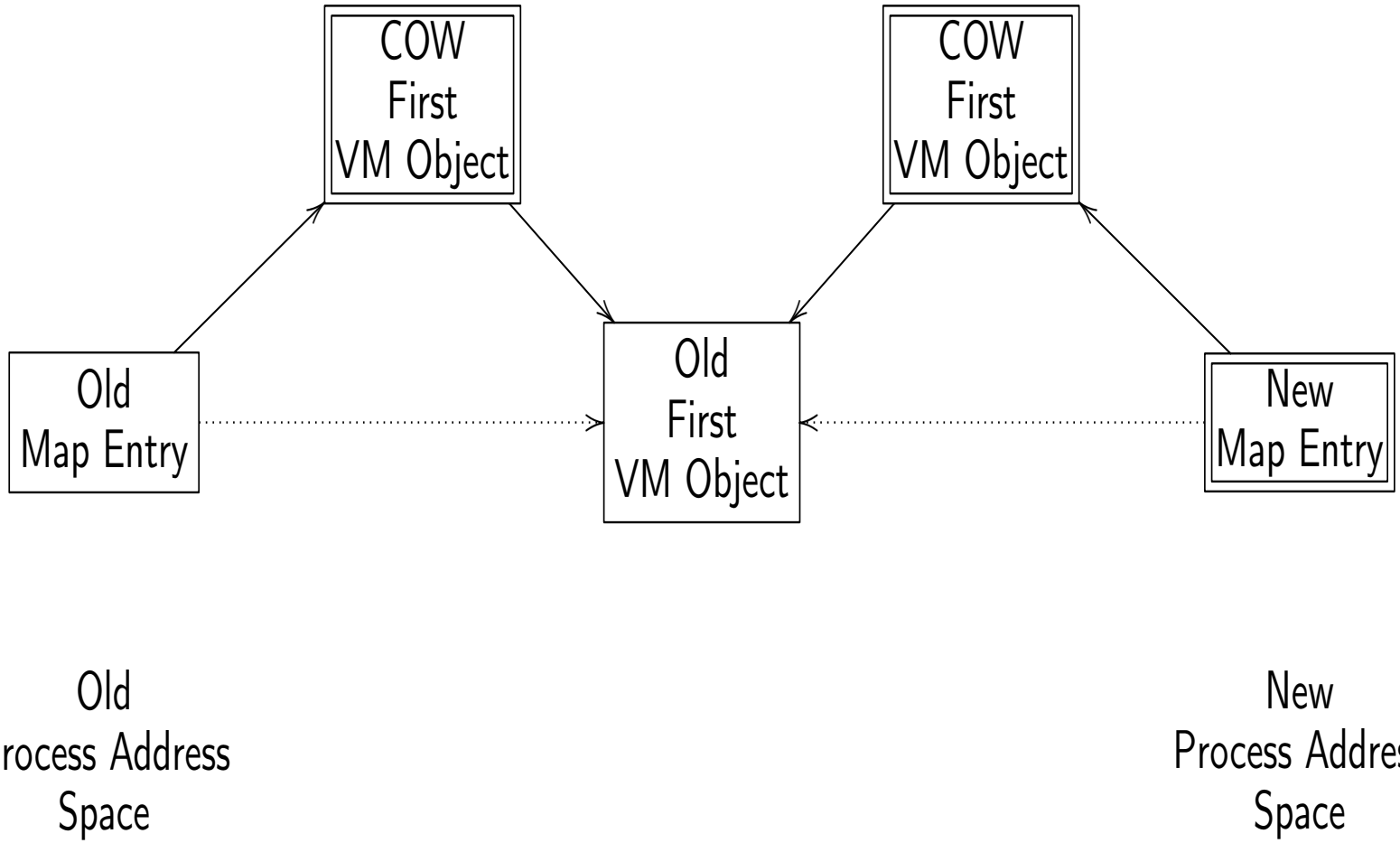
Анонимная память

OBJT_DEFAULT, OBJT_SWAP

Как выделяется

- mmap(MAP_ANON);
- fork(2)

После fork(2)


COW
First
VM Object
COW
First
VM Object
Old
Map Entry
Old
First
VM Object
New
Map Entry
Old
Process Address
Space
New
Process Address
Space

Page Fault Handler

1. Найти map entry. Проверить права доступа.
2. Найти первый объект; создать, если нет или MAP_ENTRY_NEEDS_COPY.
3. Содержит ли первый объект страницу ? Да – 7, Нет - 4.
4. Выделить свободную страницу.
5. Найти какой-нибудь объект, содержащий страницу по нужному смещению.
6. Прочитать найденную страницу, скопировать ее в выделенную.
7. Возврат из обработчика.

Swap Accounting

Когда ?

- mmap(MAP_ANON) или mmap(MAP_PRIVATE)
- fork(2)
- mprotect(2)
- ptrace(2) PT_WRITE_I (отладчики)

Disabling Overcommit

See tuning(7). Доступно, начиная с FreeBSD 8.

Основные sysctl

- vm.swap_total
- vm.swap_reserved
- vm.overcommit
  - Bit 0 – Disable swap_reserved become more then swap_total
  - Bit 1 – Enable RLIMIT_SWAP per-uid limit.
  - Bit 2 – Enable to use physical memory as swap.