Linux memory management at scale

Chris Down
Kernel, Facebook
https://chrisdown.name
Downloads

Please select the amount of RAM to download:

1GB
- 1GB CT12864AA800 Memory
- 240-pin DIMM
- DDR2 PC2-6400, CL=6

Was: $59.99 Now: FREE

☑️ Download Now

2GB
- 2GB (2 x 1GB)
- 240-pin DIMM
- DDR2 800 MHz (PC2-6400)

Was: $149.99 Now: FREE

☑️ Download Now

4GB
- 4GB (2 x 2GB)
- 240-pin DIMM
- DDR2 800 MHz (PC2-6400)

Was: $499.99 Now: FREE

☑️ Download Now
facebook

cgroupv2: Linux’s new unified control group system

Chris Down (cdown@fb.com)
Production Engineer, Web Foundation
server
```c
atomic_t mm_count;

#ifdef CONFIG_MMU
    atomic_long_t pgtables_bytes; /* PTE page table pages */
#else
    int map_count; /* number of VMAs */
#endif

spinlock_t page_table_lock; /* Protects page tables and some counters */

struct rw_semaphore mmap_sem;

struct list_head mmlist; /* List of maybe swapped mm's. These are globally strung together off init_mm.mmlist, and are protected by mmlist_lock */

unsigned long hiwater_rss; /* High-watermark of RSS usage */
unsigned long hiwater_vm; /* High-water virtual memory usage */
```

```
"include/linux/mm_types.h" 748L, 23478C
```
Memory is divided into multiple “types”: anon, cache, buffers, etc.
“Reclaimable” or “unreclaimable” is important, but not guaranteed.
RSS is kinda bullshit, sorry.
Swap isn’t about emergency memory, in fact that’s probably harmful

Instead, it increases reclaim equality and reliability of forward progress of the system

Also promotes maintaining a small positive pressure (similar to `make -j cores+1`)
- OOM killer is reactive, not proactive, based on reclaim failure
- Hotness obscured by MMU (`pte_young`), we don’t know we’re OOMing ahead of time
- Can be very, very late to the party, and sometimes go to the wrong party entirely
- kswapd reclaim: background, started when resident pages goes above a threshold
- Direct reclaim: blocks application when have no memory available to allocate frames
- Tries to reclaim the coldest pages first
- Some things might not be reclaimable. Swap can help here (bit.ly/whyswap)
“If I had more of this resource, I could probably run N% faster”

- Find bottlenecks
- Detect workload health issues before they become severe
- Used for resource allocation, load shedding, pre-OOM detection

```bash
$ cat /sys/fs/cgroup/system.slice/memory.pressure
```
```
some avg10=0.21 avg60=0.22 total=4760988587
full avg10=0.21 avg60=0.22 total=4681731696
```
Early-warning OOM detection and handling using new memory pressure metrics

Highly configurable policy/rule engine

Workload QoS and context-aware decisions
Shift to “protection” mentality

- Limits (eg. memory.{high,max}) really don’t compose well
- Prefer protection (memory.{low,min}) if possible
- Protections affect memory reclaim behaviour
- **Workload protection**: Prevent non-critical services degrading main workload
- **Host protection**: Degrade gracefully if machine cannot sustain workload
- **Usability**: Avoid introducing performance or operational costs
Base OS
  - Filesystems
  - Swap
  - Kernel tunables
  - ...

fbtax2
cgroup v2
  - Default hierarchy
  - Resource configuration

Applications
  - oomd
  - Metric exporting for cgroups
Base OS

- **btrfs as /**
  - ext4 has priority inversions
  - All metadata is annotated

- **Swap**
  - Allows memory pressure to build up gracefully
  - Usually disabled on main workload
  - btrfs swap file support to avoid tying to provisioning

- **Kernel tunables**
  - `vm.swappiness`
  - Writeback throttling
fbtax2 cgroup hierarchy: old

- **system.slice**
  - memory.high: 8G
  - memory.max: 10G
  - **hostcritical.slice**
    - web
    - **workload.slice**
      - **workload-container.slice** → HHVM
      - **workload-deps.slice**
        - Service discovery
        - Config service
  - Chef
  - sshd
  - syslog
fbtax2 cgroup hierarchy

system.slice
- io.latency: 75ms
  - hostcritical.slice
    - memory.min: 352M
      - io.latency: 50ms
        - workload.slice
          - memory.low: 17G
            - io.latency: 50ms
              - workload-container.slice
                - memory.low: max
                  - workload-deps.slice
                    - memory.low: 2.5G
                      - Service discovery
                        - Config service
  - Chef
    - sshd
      - syslog

web

HHVM
webservers: protection against memory starvation
Try it out: bit.ly/fbtax2