

facebook

Linux memory management at scale

Chris Down

Kernel, Facebook

<https://chrisdown.name>

Downloads

Please select the amount of RAM to download:

1GB



Overview

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

Was: ~~\$99.99~~ Now: **FREE**

 [Download Now](#)

2GB



Overview

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

Was: ~~\$149.99~~ Now: **FREE**

 [Download Now](#)

4GB



Overview

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

Was: ~~\$499.99~~ Now: **FREE**

 [Download Now](#)



Filmed at
QCon London 2017

Brought to you by
InfoQ

facebook

cgroupv2: Linux's new unified control group system

Chris Down (cdown@fb.com)
Production Engineer, Web Foundation

server

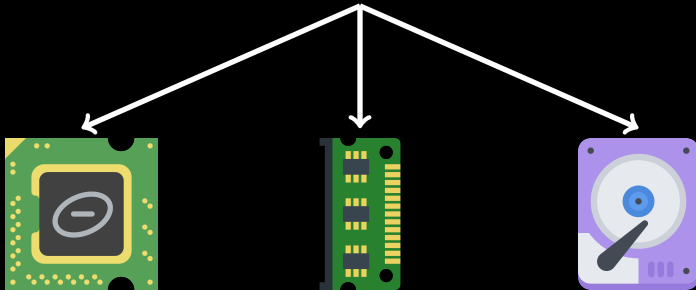




Image: Spc. Christopher Hernandez, US Military Public Domain



Image: Simon Law on Flickr, CC-BY-SA



**SRE
CON** ASIA
PACIFIC

```
10         atomic_t mm_count;
11
12     #ifdef CONFIG_MMU
13         atomic_long_t pgtables_bytes; /* PTE page table pages */
14     #endif
15     int map_count; /* number of VMAs */
16
17     spinlock_t page_table_lock; /* Protects page tables and some
18                                 * counters
19                                 */
20
21     struct rw_semaphore mmap_sem;
22
23     struct list_head mmlist; /* List of maybe swapped mm's. These
24                             * are globally strung together off
25                             * init_mm.mmlist, and are protected
26                             * by mmlist_lock
27                             */
28
29     unsigned long hiwater_rss; /* High-watermark of RSS usage */
30     unsigned long hiwater_vm; /* High-water virtual memory usage */
31
32 #include <linux/mm_types.h> 740L, 23470C
```



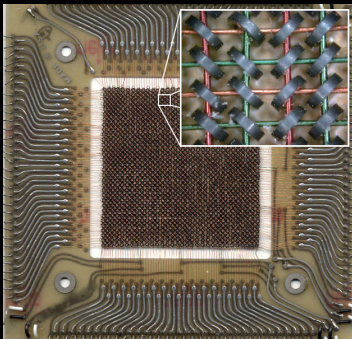


Image: Orion J on Wikimedia Commons, CC-BY

- Memory is divided in to multiple “types”: anon, cache, buffers, etc
- “Reclaimable” or “unreclaimable” is important, but not guaranteed
- RSS is kinda bullshit, sorry

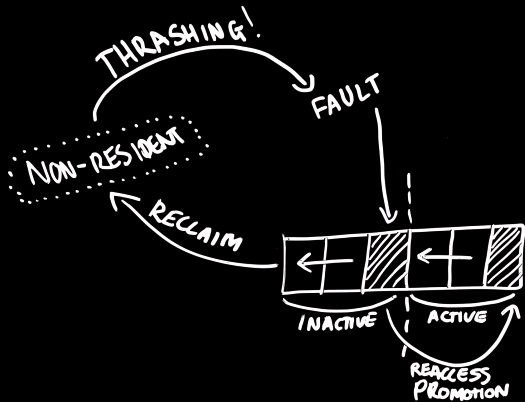


bit.ly/whyswap

- 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)



psi

“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

```
$ 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
```



bit.ly/fboomd

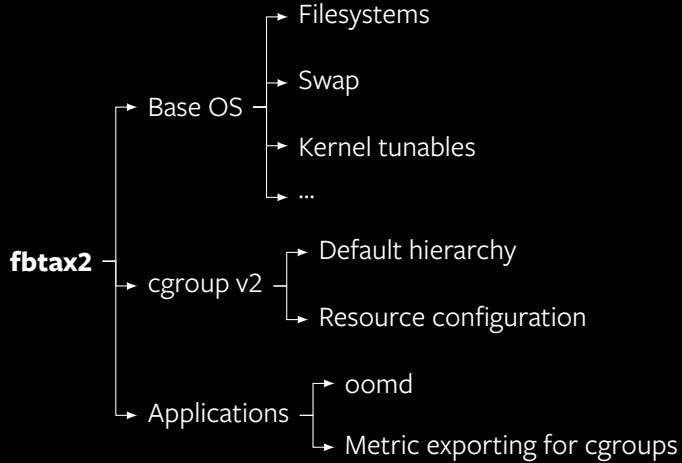
- Early-warning OOM detection and handling using new memory pressure metrics
- Highly configurable policy/rule engine
- Workload QoS and context-aware decisions

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

fbtax2

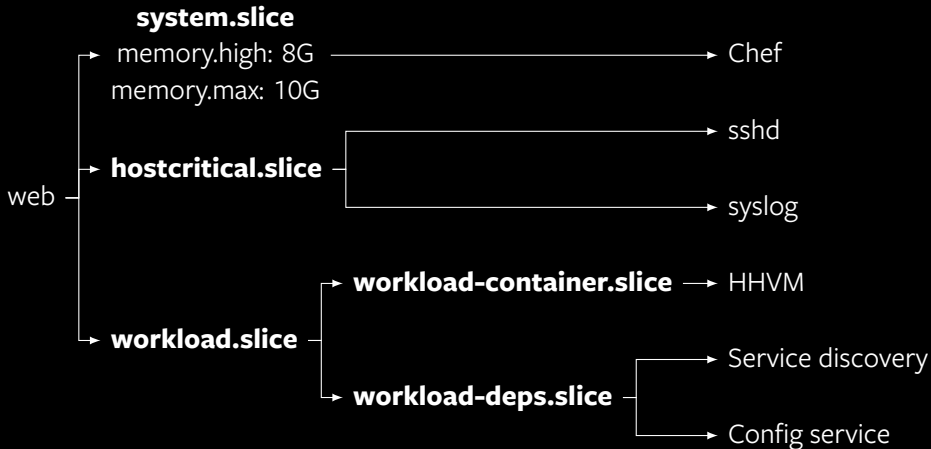
- **Workload protection:** Prevent non-critical services degrading main workload
- **Host protection:** Degrade gracefully if machine cannot sustain workload
- **Usability:** Avoid introducing performance or operational costs



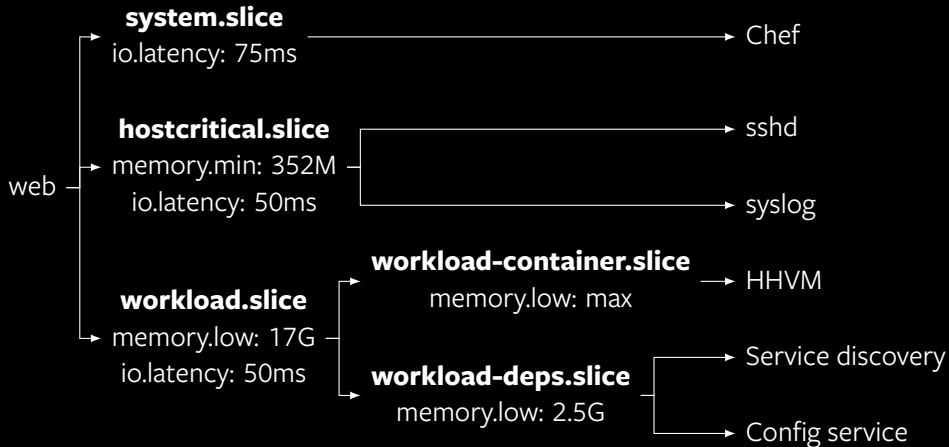
Base OS

- btrfs as /
 - ext4 has priority inversions
 - All metadata is annotated
- Swap
 - Yes, you really still want it (bit.ly/whyswap)
 - 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

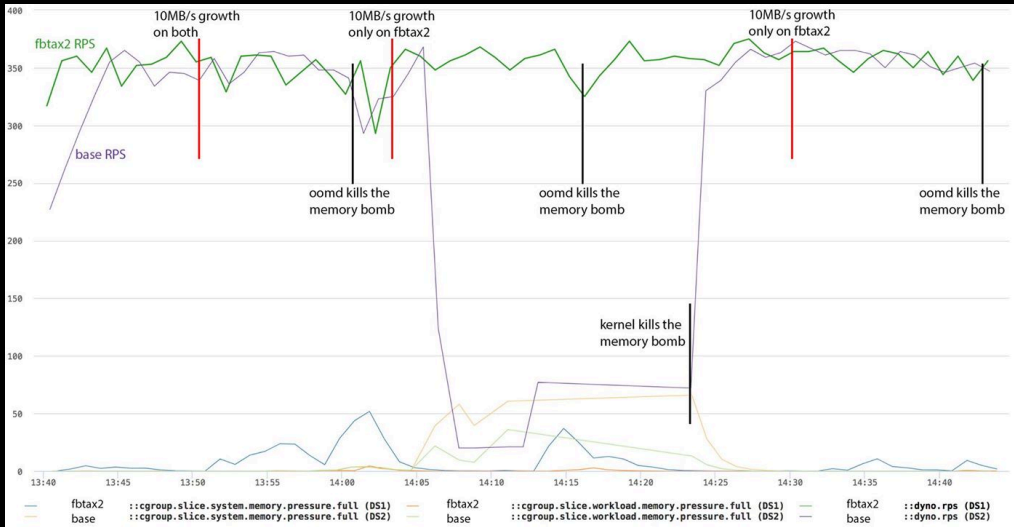
fbtax2 cgroup hierarchy: old



fbtax2 cgroup hierarchy



webservers: protection against memory starvation



Try it out: bit.ly/fbtax2

facebook