Subject: [PATCH 18/23] slub: charge allocation to a memcg
Posted by Glauber Costa on Sun, 22 Apr 2012 23:53:35 GMT
View Forum Message <> Reply to Message

This patch charges allocation of a slab object to a particular
memcg.

The cache is selected with mem_cgroup_get_kmem_cache(),
which is the biggest overhead we pay here, because
it happens at all allocations. However, other than forcing
a function call, this function is not very expensive, and
try to return as soon as we realize we are not a memcg cache.

The charge/uncharge functions are heavier, but are only called
for new page allocations.

The kmalloc_no_account variant is patched so the base
function is used and we don't even try to do cache
selection.

Signed-off-by: Glauber Costa <glommer@parallels.com>
CC: Christoph Lameter <cl@linux.com>
CC: Pekka Enberg <penberg@cs.helsinki.fi>
CC: Michal Hocko <mhocko@suse.cz>
CC: Kamezawa Hiroyuki <kamezawa.hiroyu@jp.fujitsu.com>
CC: Johannes Weiner <hannes@cmpxchg.org>
CC: Suleiman Souhlal <suleiman@google.com>
---
 include/linux/slub_def.h | 32 ++++++++++--
 mm/slub.c | 124 ++-----
 2 files changed, 138 insertions(+), 18 deletions(-)

diff --git a/include/linux/slub_def.h b/include/linux/slub_def.h
index 9a8000a..e75efcb 100644
--- a/include/linux/slub_def.h
+++ b/include/linux/slub_def.h
@@ -13,6 +13,7 @@
 #include <linux/kobject.h>

 #include <linux/kmemleak.h>
+#include <linux/memcontrol.h>

 enum stat_item {
  ALLOC_FASTPATH, /* Allocation from cpu slab */
@@ -210,14 +211,21 @@ static __always_inline int kmalloc_index(size_t size)
  * This ought to end up with a global pointer to the right cache
  * in kmalloc_caches.
  */

```
-static __always_inline struct kmem_cache *kmalloc_slab(size_t size)
+static __always_inline struct kmem_cache *kmalloc_slab(gfp_t flags, size_t size)
 {
+ struct kmem_cache *s;
  int index = kmalloc_index(size);

  if (index == 0)
   return NULL;

- return kmalloc_caches[index];
+ s = kmalloc_caches[index];
+
+ rcu_read_lock();
+ s = mem_cgroup_get_kmem_cache(s, flags);
+ rcu_read_unlock();
+
+ return s;
 }

 void *kmem_cache_alloc(struct kmem_cache *, gfp_t);
@@ -225,13 +233,27 @@ void *kmalloc_no_account(size_t size, gfp_t);
 void *__kmalloc(size_t size, gfp_t flags);

 static __always_inline void *
-kmalloc_order(size_t size, gfp_t flags, unsigned int order)
+kmalloc_order_base(size_t size, gfp_t flags, unsigned int order)
 {
  void *ret = (void *) __get_free_pages(flags | __GFP_COMP, order);
  kmemleak_alloc(ret, size, 1, flags);
  return ret;
 }

+static __always_inline void *
+kmalloc_order(size_t size, gfp_t flags, unsigned int order)
+{
+ void *ret = NULL;
+
+ if (!mem_cgroup_charge_kmem(flags, size))
+  return NULL;
+
+ ret = kmalloc_order_base(size, flags, order);
+ if (!ret)
+  mem_cgroup_uncharge_kmem((1 << order) << PAGE_SHIFT);
+ return ret;
+}
+
 /**
  * Calling this on allocated memory will check that the memory
```

```
  * is expected to be in use, and print warnings if not.
@@ -276,7 +298,7 @@ static __always_inline void *kmalloc(size_t size, gfp_t flags)
    return kmalloc_large(size, flags);

   if (!(flags & SLUB_DMA)) {
-   struct kmem_cache *s = kmalloc_slab(size);
+   struct kmem_cache *s = kmalloc_slab(flags, size);

    if (!s)
     return ZERO_SIZE_PTR;
@@ -309,7 +331,7 @@ static __always_inline void *kmalloc_node(size_t size, gfp_t flags, int
node)
 {
  if (__builtin_constant_p(size) &&
   size <= SLUB_MAX_SIZE && !(flags & SLUB_DMA)) {
-   struct kmem_cache *s = kmalloc_slab(size);
+   struct kmem_cache *s = kmalloc_slab(flags, size);

   if (!s)
    return ZERO_SIZE_PTR;
diff --git a/mm/slub.c b/mm/slub.c
index d754b06..9b22139 100644
--- a/mm/slub.c
+++ b/mm/slub.c
@@ -1283,11 +1283,17 @@ static inline struct page *alloc_slab_page(gfp_t flags, int node,
   return alloc_pages_exact_node(node, flags, order);
 }

+static inline unsigned long size_in_bytes(unsigned int order)
+{
+ return (1 << order) << PAGE_SHIFT;
+}
+
 static struct page *allocate_slab(struct kmem_cache *s, gfp_t flags, int node)
 {
- struct page *page;
+ struct page *page = NULL;
  struct kmem_cache_order_objects oo = s->oo;
  gfp_t alloc_gfp;
+ unsigned int memcg_allowed = oo_order(oo);

  flags &= gfp_allowed_mask;

@@ -1296,13 +1302,29 @@ static struct page *allocate_slab(struct kmem_cache *s, gfp_t flags,
int node)

  flags |= s->allocflags;
```

```
- /*
-  * Let the initial higher-order allocation fail under memory pressure
-  * so we fall-back to the minimum order allocation.
-  */
- alloc_gfp = (flags | __GFP_NOWARN | __GFP_NORETRY) & ~__GFP_NOFAIL;
+ memcg_allowed = oo_order(oo);
+ if (!mem_cgroup_charge_slab(s, flags, size_in_bytes(memcg_allowed))) {
+
+  memcg_allowed = oo_order(s->min);
+  if (!mem_cgroup_charge_slab(s, flags,
+         size_in_bytes(memcg_allowed))) {
+   if (flags & __GFP_WAIT)
+    local_irq_disable();
+   return NULL;
+  }
+ }
+
+ if (memcg_allowed == oo_order(oo)) {
+  /*
+   * Let the initial higher-order allocation fail under memory
+   * pressure so we fall-back to the minimum order allocation.
+   */
+  alloc_gfp = (flags | __GFP_NOWARN | __GFP_NORETRY) &
+        ~__GFP_NOFAIL;
+
+  page = alloc_slab_page(alloc_gfp, node, oo);
+ }

- page = alloc_slab_page(alloc_gfp, node, oo);
  if (unlikely(!page)) {
   oo = s->min;
   /*
@@ -1313,13 +1335,23 @@ static struct page *allocate_slab(struct kmem_cache *s, gfp_t flags,
int node)

   if (page)
    stat(s, ORDER_FALLBACK);
+  /*
+   * We reserved more than we used, time to give it back
+   */
+  if (page && memcg_allowed != oo_order(oo)) {
+   unsigned long delta;
+   delta = memcg_allowed - oo_order(oo);
+   mem_cgroup_uncharge_slab(s, size_in_bytes(delta));
+  }
  }

  if (flags & __GFP_WAIT)
```

```
   local_irq_disable();

- if (!page)
+ if (!page) {
+  mem_cgroup_uncharge_slab(s, size_in_bytes(memcg_allowed));
   return NULL;
+ }

  if (kmemcheck_enabled
   && !(s->flags & (SLAB_NOTRACK | DEBUG_DEFAULT_FLAGS))) {
@@ -1393,6 +1425,24 @@ out:
  return page;
 }

+#ifdef CONFIG_CGROUP_MEM_RES_CTLR_KMEM
+static unsigned long slab_nr_pages(struct kmem_cache *s)
+{
+ int node;
+ unsigned long nr_slabs = 0;
+
+ for_each_online_node(node) {
+  struct kmem_cache_node *n = get_node(s, node);
+
+  if (!n)
+   continue;
+  nr_slabs += atomic_long_read(&n->nr_slabs);
+ }
+
+ return nr_slabs << oo_order(s->oo);
+}
+#endif
+
 static void __free_slab(struct kmem_cache *s, struct page *page)
 {
  int order = compound_order(page);
@@ -1419,6 +1469,12 @@ static void __free_slab(struct kmem_cache *s, struct page *page)
  if (current->reclaim_state)
   current->reclaim_state->reclaimed_slab += pages;
  __free_pages(page, order);
+
+ mem_cgroup_uncharge_slab(s, (1 << order) << PAGE_SHIFT);
+#ifdef CONFIG_CGROUP_MEM_RES_CTLR_KMEM
+ if (s->memcg_params.memcg && (slab_nr_pages(s) == 0))
+  mem_cgroup_destroy_cache(s);
+#endif
 }

 #define need_reserve_slab_rcu      \
```

```
@@ -2300,8 +2356,9 @@ new_slab:
  *
  * Otherwise we can simply pick the next object from the lockless free list.
  */
-static __always_inline void *slab_alloc(struct kmem_cache *s,
-  gfp_t gfpflags, int node, unsigned long addr)
+static __always_inline void *slab_alloc_base(struct kmem_cache *s,
+          gfp_t gfpflags, int node,
+          unsigned long addr)
 {
  void **object;
  struct kmem_cache_cpu *c;
@@ -2369,6 +2426,24 @@ redo:
  return object;
 }

+static __always_inline void *slab_alloc(struct kmem_cache *s,
+  gfp_t gfpflags, int node, unsigned long addr)
+{
+
+ if (slab_pre_alloc_hook(s, gfpflags))
+  return NULL;
+
+ if (in_interrupt() || (current == NULL) || (gfpflags & __GFP_NOFAIL))
+  goto kernel_alloc;
+
+ rcu_read_lock();
+ s = mem_cgroup_get_kmem_cache(s, gfpflags);
+ rcu_read_unlock();
+
+kernel_alloc:
+ return slab_alloc_base(s, gfpflags, node, addr);
+}
+
 void *kmem_cache_alloc(struct kmem_cache *s, gfp_t gfpflags)
 {
  void *ret = slab_alloc(s, gfpflags, NUMA_NO_NODE, _RET_IP_);
@@ -3194,6 +3269,13 @@ void kmem_cache_destroy(struct kmem_cache *s)
  s->refcount--;
  if (!s->refcount) {
   list_del(&s->list);
+#ifdef CONFIG_CGROUP_MEM_RES_CTLR_KMEM
+  /* Not a memcg cache */
+  if (s->memcg_params.id != -1) {
+   mem_cgroup_release_cache(s);
+   mem_cgroup_flush_cache_create_queue();
+  }
+#endif
```

```
   up_write(&slub_lock);
   if (kmem_cache_close(s)) {
    printk(KERN_ERR "SLUB %s: %s called for cache that "
@@ -3273,6 +3355,7 @@ static struct kmem_cache *__init create_kmalloc_cache(const char
*name,
   goto panic;

  list_add(&s->list, &slab_caches);
+ mem_cgroup_register_cache(NULL, s);
  return s;

 panic:
@@ -3364,15 +3447,21 @@ void *kmalloc_no_account(size_t size, gfp_t flags)
  struct kmem_cache *s;
  void *ret;

- if (unlikely(size > SLUB_MAX_SIZE))
-  return kmalloc_large(size, flags);
+ if (unlikely(size > SLUB_MAX_SIZE)) {
+  unsigned int order = get_order(size);
+  ret = kmalloc_order_base(size, flags, order);
+#ifdef CONFIG_TRACING
+  trace_kmalloc(_RET_IP_, ret, size, PAGE_SIZE << order, flags);
+#endif
+  return ret;
+ }

  s = get_slab(size, flags);

  if (unlikely(ZERO_OR_NULL_PTR(s)))
   return s;

- ret = slab_alloc(s, flags, NUMA_NO_NODE, _RET_IP_);
+ ret = slab_alloc_base(s, flags, NUMA_NO_NODE, _RET_IP_);

  trace_kmalloc(_RET_IP_, ret, size, s->size, flags);

@@ -3387,10 +3476,17 @@ static void *kmalloc_large_node(size_t size, gfp_t flags, int node)
  void *ptr = NULL;

  flags |= __GFP_COMP | __GFP_NOTRACK;
+
+ if (!mem_cgroup_charge_kmem(flags, size))
+  goto out;
+
  page = alloc_pages_node(node, flags, get_order(size));
  if (page)
   ptr = page_address(page);
```

```
+ else
+  mem_cgroup_uncharge_kmem(size);

+out:
  kmemleak_alloc(ptr, size, 1, flags);
  return ptr;
 }
@@ -3938,8 +4034,10 @@ static struct kmem_cache *find_mergeable(struct mem_cgroup
*memcg, size_t size,
   if (s->size - size >= sizeof(void *))
    continue;

+#ifdef CONFIG_CGROUP_MEM_RES_CTLR_KMEM
   if (memcg && s->memcg_params.memcg != memcg)
    continue;
+#endif

   return s;
  }
--
1.7.7.6
```