Subject: [PATCH 0/2] dm-ioband v0.0.3: The I/O bandwidth controller: Introduction
Posted by Ryo Tsuruta on Tue, 05 Feb 2008 10:17:35 GMT
View Forum Message <> Reply to Message

Hi everyone,

This is dm-ioband version 0.0.3 release.

Dm-ioband is an I/O bandwidth controller implemented as a device-mapper driver,
which gives specified bandwidth to each job running on the same physical
device.

Changes since 0.0.2 (23rd January):
- Ported to linux-2.6.24.
- Rename the name of this device-mapper device as "ioband."
- The output format of "dmsetup table" can be recognized by "dmsetup
  create/load/reload."
- Add a new feature to block processes requesting BIOs in case the number
  of the uncompleted BIOs becomes too large.
- Gid can be used as an ioband group type.
- Support "dmsetup suspend --noflush."
- Fix the problem that "dmsetup detach" doesn't wait the completion of
  all the BIOs in the detaching group.
- Fix the problem that "dmsetup message attach" can't accept ID 0.
- Code cleanups. Use prefix "dm_" for the global symbols.
- Add examples in the document.

Thanks,
Ryo Tsuruta

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

Subject: [PATCH 1/2] dm-ioband v0.0.3: The I/O bandwidth controller: Source code patch
Posted by Ryo Tsuruta on Tue, 05 Feb 2008 10:18:46 GMT
View Forum Message <> Reply to Message

Here is the patch of dm-ioband.

Based on 2.6.24
Signed-off-by: Ryo Tsuruta <ryov@valinux.co.jp>
Signed-off-by: Hirokazu Takahashi <taka@valinux.co.jp>

diff -uprN linux-2.6.24.orig/drivers/md/Kconfig linux-2.6.24/drivers/md/Kconfig
--- linux-2.6.24.orig/drivers/md/Kconfig 2008-01-25 07:58:37.000000000 +0900

```
+++ linux-2.6.24/drivers/md/Kconfig 2008-02-05 19:09:41.000000000 +0900
@@ -288,4 +288,17 @@ config DM_UEVENT
  ---help---
  Generate udev events for DM events.

+config DM_IOBAND
+ tristate "I/O bandwidth control (EXPERIMENTAL)"
+ depends on BLK_DEV_DM && EXPERIMENTAL
+ ---help---
+ This device-mapper target allows to define how the
+ available bandwith of a storage device should be
+ shared between processes, cgroups, the partitions or the LUNs.
+
+ Information on how to use dm-ioband is available in:
+    <file:Documentation/device-mapper/ioband.txt>.
+
+ If unsure, say N.
+
 endif # MD
diff -uprN linux-2.6.24.orig/drivers/md/Makefile linux-2.6.24/drivers/md/Makefile
--- linux-2.6.24.orig/drivers/md/Makefile 2008-01-25 07:58:37.000000000 +0900
+++ linux-2.6.24/drivers/md/Makefile 2008-02-05 19:09:41.000000000 +0900
@@ -9,6 +9,7 @@ dm-snapshot-objs := dm-snap.o dm-excepti
 dm-mirror-objs := dm-log.o dm-raid1.o
 dm-rdac-objs := dm-mpath-rdac.o
 dm-hp-sw-objs := dm-mpath-hp-sw.o
+dm-ioband-objs := dm-ioband-ctl.o dm-ioband-policy.o dm-ioband-type.o
 md-mod-objs     := md.o bitmap.o
 raid456-objs := raid5.o raid6algos.o raid6recov.o raid6tables.o \
      raid6int1.o raid6int2.o raid6int4.o \
@@ -41,6 +42,7 @@ obj-$(CONFIG_DM_MULTIPATH_RDAC) += dm-rd
 obj-$(CONFIG_DM_SNAPSHOT) += dm-snapshot.o
 obj-$(CONFIG_DM_MIRROR)  += dm-mirror.o
 obj-$(CONFIG_DM_ZERO)  += dm-zero.o
+obj-$(CONFIG_DM_IOBAND)  += dm-ioband.o

 quiet_cmd_unroll = UNROLL  $@
       cmd_unroll = $(PERL) $(srctree)/$(src)/unroll.pl $(UNROLL) \
diff -uprN linux-2.6.24.orig/drivers/md/dm-ioband-ctl.c linux-2.6.24/drivers/md/dm-ioband-ctl.c
--- linux-2.6.24.orig/drivers/md/dm-ioband-ctl.c 1970-01-01 09:00:00.000000000 +0900
+++ linux-2.6.24/drivers/md/dm-ioband-ctl.c 2008-02-05 19:09:41.000000000 +0900
@@ -0,0 +1,947 @@
+/*
+ * Copyright (C) 2008 VA Linux Systems Japan K.K.
+ * Authors: Hirokazu Takahashi <taka@valinux.co.jp>
+ *          Ryo Tsuruta <ryov@valinux.co.jp>
+ *
+ *  I/O bandwidth control
```

```
+ *
+ * This file is released under the GPL.
+ */
+#include <linux/module.h>
+#include <linux/init.h>
+#include <linux/bio.h>
+#include <linux/slab.h>
+#include <linux/workqueue.h>
+#include <linux/raid/md.h>
+#include "dm.h"
+#include "dm-bio-list.h"
+#include "dm-ioband.h"
+
+#define DM_MSG_PREFIX "ioband"
+
+static LIST_HEAD(ioband_device_list);
+/* to protect ioband_device_list */
+static DEFINE_SPINLOCK(ioband_devicelist_lock);
+
+#if LINUX_VERSION_CODE < KERNEL_VERSION(2,6,24)
+static void ioband_conduct(void *);
+#else
+static void ioband_conduct(struct work_struct *);
+#endif
+static void ioband_hold_bio(struct ioband_group *, struct bio *);
+static struct bio *ioband_pop_bio(struct ioband_group *);
+static int ioband_set_param(struct ioband_group *, char *, char *);
+static int ioband_group_attach(struct ioband_group *, int);
+
+int ioband_debug; /* just for debugging */
+
+static void policy_init(struct ioband_device *dp, char *name)
+{
+ struct policy_type *p;
+ for (p = dm_ioband_policy_type; (p->p_name); p++) {
+  if (!strcmp(name, p->p_name))
+   break;
+ }
+
+ p->p_policy_init(dp);
+ if (!dp->g_hold_bio)
+  dp->g_hold_bio = ioband_hold_bio;
+ if (!dp->g_pop_bio)
+  dp->g_pop_bio = ioband_pop_bio;
+}
+
+static struct ioband_device *alloc_ioband_device(int devgroup_id, char *name,
+    int io_throttle, int io_limit)
```

```
+
+{
+ struct ioband_device *dp = NULL;
+ struct ioband_device *p;
+ struct ioband_device *new;
+ unsigned long flags;
+
+ new = kzalloc(sizeof(struct ioband_device), GFP_KERNEL);
+ if (!new)
+  goto try_to_find;
+
+ /*
+  * Prepare its own workqueue as generic_make_request() may potentially
+  * block the workqueue when submitting BIOs.
+  */
+ new->g_ioband_wq = create_workqueue("kioband");
+ if (!new->g_ioband_wq) {
+  kfree(new);
+  new = NULL;
+  goto try_to_find;
+ }
+
+#if LINUX_VERSION_CODE < KERNEL_VERSION(2,6,24)
+ INIT_WORK(&new->g_conductor, ioband_conduct, new);
+#else
+ INIT_WORK(&new->g_conductor, ioband_conduct);
+#endif
+ INIT_LIST_HEAD(&new->g_groups);
+ INIT_LIST_HEAD(&new->g_list);
+ spin_lock_init(&new->g_lock);
+ new->g_devgroup = devgroup_id;
+ new->g_io_throttle = io_throttle;
+ new->g_io_limit = io_limit;
+ new->g_plug_bio = 0;
+ new->g_issued = 0;
+ new->g_blocked = 0;
+ new->g_ref = 0;
+ new->g_flags = 0;
+ memset(new->g_name, 0, sizeof(new->g_name));
+ new->g_hold_bio = NULL;
+ new->g_pop_bio = NULL;
+ init_waitqueue_head(&new->g_waitq);
+
+try_to_find:
+ spin_lock_irqsave(&ioband_devicelist_lock, flags);
+ list_for_each_entry(p, &ioband_device_list, g_list) {
+  if (p->g_devgroup == devgroup_id) {
+   dp = p;
```

```
+   break;
+  }
+ }
+ if (!dp && (new)) {
+  policy_init(new, name);
+  dp = new;
+  new = NULL;
+  list_add_tail(&dp->g_list, &ioband_device_list);
+ }
+ spin_unlock_irqrestore(&ioband_devicelist_lock, flags);
+
+ if (new) {
+  destroy_workqueue(new->g_ioband_wq);
+  kfree(new);
+ }
+
+ return dp;
+}
+
+static inline void release_ioband_device(struct ioband_device *dp)
+{
+ unsigned long flags;
+
+ spin_lock_irqsave(&ioband_devicelist_lock, flags);
+ if (!list_empty(&dp->g_groups)) {
+  spin_unlock_irqrestore(&ioband_devicelist_lock, flags);
+  return;
+ }
+ list_del(&dp->g_list);
+ spin_unlock_irqrestore(&ioband_devicelist_lock, flags);
+ destroy_workqueue(dp->g_ioband_wq);
+ kfree(dp);
+}
+
+static struct ioband_group *ioband_group_find(struct ioband_group *head,int id)
+{
+ struct ioband_group *p;
+ struct ioband_group *gp = NULL;
+
+ list_for_each_entry(p, &head->c_group_list, c_group_list) {
+  if (p->c_id == id || id == IOBAND_ID_ANY)
+   gp = p;
+ }
+ return gp;
+}
+
+static int ioband_group_init(struct ioband_group *gp,
+  struct ioband_group *head, struct ioband_device *dp, int id)
```

```
+{
+ unsigned long flags;
+
+ INIT_LIST_HEAD(&gp->c_list);
+ bio_list_init(&gp->c_blocked_bios);
+ gp->c_id = id; /* should be verified */
+ gp->c_blocked = 0;
+ memset(gp->c_stat, 0, sizeof(gp->c_stat));
+ init_waitqueue_head(&gp->c_waitq);
+ gp->c_flags = 0;
+
+ INIT_LIST_HEAD(&gp->c_group_list);
+
+ gp->c_banddev = dp;
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ if (head && ioband_group_find(head, id)) {
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  DMWARN("ioband_group: id=%d already exists.", id);
+  return -EEXIST;
+ }
+ dp->g_ref++;
+ list_add_tail(&gp->c_list, &dp->g_groups);
+
+ dp->g_group_ctr(gp);
+
+ if (head) {
+  list_add_tail(&gp->c_group_list, &head->c_group_list);
+  gp->c_dev = head->c_dev;
+  gp->c_target = head->c_target;
+ }
+
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+
+ return 0;
+}
+
+static inline void ioband_group_release(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ list_del(&gp->c_list);
+ list_del(&gp->c_group_list);
+ dp->g_ref--;
+ dp->g_group_dtr(gp);
+ kfree(gp);
+}
+
```

```
+static void ioband_group_destroy_all(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct ioband_group *group;
+ unsigned long flags;
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ while ((group = ioband_group_find(gp, IOBAND_ID_ANY)))
+  ioband_group_release(group);
+ ioband_group_release(gp);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+}
+
+static void ioband_group_stop(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ unsigned long flags;
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ set_group_down(gp);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+ queue_work(dp->g_ioband_wq, &dp->g_conductor);
+ flush_workqueue(dp->g_ioband_wq);
+}
+
+static void ioband_group_stop_all(struct ioband_group *head, int suspend)
+{
+ struct ioband_device *dp = head->c_banddev;
+ struct ioband_group *p;
+ unsigned long flags;
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ list_for_each_entry(p, &head->c_group_list, c_group_list) {
+  set_group_down(p);
+  if (suspend) {
+   set_group_suspended(p);
+   dprintk(KERN_ERR "ioband suspend: gp(%p)\n", p);
+  }
+
+ }
+ set_group_down(head);
+ if (suspend) {
+  set_group_suspended(head);
+  dprintk(KERN_ERR "ioband suspend: gp(%p)\n", head);
+ }
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+ queue_work(dp->g_ioband_wq, &dp->g_conductor);
+ flush_workqueue(dp->g_ioband_wq);
```

```
+}
+
+static void ioband_group_resume_all(struct ioband_group *head)
+{
+ struct ioband_device *dp = head->c_banddev;
+ struct ioband_group *p;
+ unsigned long flags;
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ list_for_each_entry(p, &head->c_group_list, c_group_list) {
+  clear_group_down(p);
+  clear_group_suspended(p);
+  dprintk(KERN_ERR "ioband resume: gp(%p)\n", p);
+ }
+ clear_group_down(head);
+ clear_group_suspended(head);
+ dprintk(KERN_ERR "ioband resume: gp(%p)\n", head);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+}
+
+/*
+ * Create a new band device:
+ *   parameters:  <device> <device-group-id> [<io_throttle>] [<io_limit>]
+ */
+static int ioband_ctr(struct dm_target *ti, unsigned int argc, char **argv)
+{
+ struct ioband_group *gp;
+ struct ioband_device *dp;
+ int io_throttle = DEFAULT_IO_THROTTLE;
+ int io_limit = DEFAULT_IO_LIMIT;
+ int devgroup_id;
+ int val;
+ int r = 0;
+
+ if (argc < 2) {
+  ti->error = "Requires 2 or more arguments";
+  return -EINVAL;
+ }
+
+ gp = kzalloc(sizeof(struct ioband_group), GFP_KERNEL);
+ if (!gp) {
+  ti->error = "Cannot allocate memory for bandgroup";
+  return -ENOMEM;
+ }
+
+ val = simple_strtol(argv[1], NULL, 0);
+ if (val < 0) {
+  ti->error = "Device Group ID # is too large";
```

```
+  r = -EINVAL;
+  goto error;
+ }
+ devgroup_id = val;
+ dprintk(KERN_ERR "ioband_ctr device group id:%d\n", val);
+
+ if (argc >= 3) {
+  val = simple_strtol(argv[2], NULL, 0);
+  if (val > 0)
+   io_throttle = val;
+  dprintk(KERN_ERR "ioband_ctr ioqueue_low:%d\n", io_throttle);
+ }
+ if (argc >= 4) {
+  val = simple_strtol(argv[3], NULL, 0);
+  if (val > 0)
+   io_limit = val;
+  dprintk(KERN_ERR "ioband_ctr ioqueue_high:%d\n", io_limit);
+ }
+ if (io_limit < io_throttle)
+  io_limit = io_throttle;
+
+ if (dm_get_device(ti, argv[0], 0, ti->len,
+     dm_table_get_mode(ti->table), &gp->c_dev)) {
+  ti->error = "band: device lookup failed";
+  r = -EINVAL;
+  goto error;
+ }
+
+ dp = alloc_ioband_device(devgroup_id, "default",io_throttle, io_limit);
+ if (!dp) {
+  ti->error = "Cannot allocate memory for banddevice";
+  r = -ENOMEM;
+  goto error2;
+ }
+
+ ioband_group_init(gp, NULL, dp, IOBAND_ID_ANY);
+ gp->c_getid = dm_ioband_group_type[0].t_getid;
+
+ ti->private = gp;
+
+ return 0;
+
+error2:
+ dm_put_device(ti, gp->c_dev);
+error:
+ kfree(gp);
+ return r;
+}
```

```
+
+static void ioband_dtr(struct dm_target *ti)
+{
+ struct ioband_group *gp = ti->private;
+ struct ioband_device *dp = gp->c_banddev;
+
+ ioband_group_stop_all(gp, 0);
+ dm_put_device(ti, gp->c_dev);
+ ioband_group_destroy_all(gp);
+ release_ioband_device(dp);
+}
+
+static void ioband_hold_bio(struct ioband_group *gp, struct bio *bio)
+{
+ /* Todo: The list should be split into a read list and a write list */
+ bio_list_add(&gp->c_blocked_bios, bio);
+}
+
+static struct bio *ioband_pop_bio(struct ioband_group *gp)
+{
+ return bio_list_pop(&gp->c_blocked_bios);
+}
+
+static inline void resume_to_accept_bios(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ if (is_device_blocked(dp) && dp->g_blocked < dp->g_io_limit) {
+  clear_device_blocked(dp);
+  wake_up_all(&dp->g_waitq);
+ }
+ if (is_group_blocked(gp)) {
+  clear_group_blocked(gp);
+  wake_up_all(&gp->c_waitq);
+ }
+}
+
+static inline int device_should_block(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ if (is_group_down(gp))
+  return 0;
+ if (is_device_blocked(dp))
+  return 1;
+ if (dp->g_blocked >= dp->g_io_limit) {
+  set_device_blocked(dp);
+  return 1;
```

```
+ }
+ return 0;
+}
+
+static inline int group_should_block(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ if (is_group_down(gp))
+  return 0;
+ if (is_group_blocked(gp))
+  return 1;
+ if (dp->g_should_block(gp)) {
+  set_group_blocked(gp);
+  return 1;
+ }
+ return 0;
+}
+
+static inline void do_nothing(void) {}
+
+static inline void prevent_burst_bios(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ if (!current->mm) {
+  /*
+   * Kernel threads shouldn't be blocked easily since each of
+   * them may handle BIOs for several groups on several
+   * partitions.
+   */
+  wait_event_lock_irq(dp->g_waitq, !device_should_block(gp),
+      dp->g_lock, do_nothing());
+ } else {
+  wait_event_lock_irq(gp->c_waitq, !group_should_block(gp),
+      dp->g_lock, do_nothing());
+ }
+}
+
+static inline int should_pushback_bio(struct ioband_group *gp)
+{
+ return is_group_suspended(gp) && dm_noflush_suspending(gp->c_target);
+}
+
+static inline void prepare_to_issue(struct ioband_group *gp, struct bio *bio)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
```

```
+ dp->g_prepare_bio(gp, bio);
+ dp->g_issued++;
+ if (dp->g_issued >= dp->g_io_limit)
+  dp->g_plug_bio = 1;
+}
+
+static inline int room_for_bio(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ return !dp->g_plug_bio || is_group_down(gp);
+}
+
+static inline void hold_bio(struct ioband_group *gp, struct bio *bio)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ dp->g_blocked++;
+ gp->c_blocked++;
+ gp->c_stat[bio_data_dir(bio)].deferred++;
+ dp->g_hold_bio(gp, bio);
+}
+
+static inline int release_bios(struct ioband_group *gp,
+   struct bio_list *issue_list, struct bio_list *pushback_list)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct bio *bio;
+
+ while (dp->g_can_submit(gp) && gp->c_blocked) {
+  if (!room_for_bio(gp))
+   return 1;
+  bio = dp->g_pop_bio(gp);
+  if (!bio)
+   return 0;
+  dp->g_blocked--;
+  gp->c_blocked--;
+  if (!gp->c_blocked)
+   resume_to_accept_bios(gp);
+  prepare_to_issue(gp, bio);
+  if (is_group_suspended(gp))
+   bio_list_add(pushback_list, bio);
+  else
+   bio_list_add(issue_list, bio);
+ }
+
+ return 0;
+}
```

```
+
+static inline struct ioband_group *ioband_group_get(
+    struct ioband_group *head, struct bio *bio)
+{
+ struct ioband_group *gp;
+
+ if (!head->c_getid)
+  return head;
+
+ gp = ioband_group_find(head, head->c_getid(bio));
+
+ if (!gp)
+  gp = head;
+ return gp;
+}
+
+/*
+ * Start to control the bandwidth once the number of uncompleted BIOs
+ * exceeds the value of "io_throttle".
+ */
+static int ioband_map(struct dm_target *ti, struct bio *bio,
+       union map_info *map_context)
+{
+ struct ioband_group *gp = ti->private;
+ struct ioband_group_stat *bws;
+ struct ioband_device *dp = gp->c_banddev;
+ unsigned long flags;
+
+#if 0 /* not supported yet */
+ if (bio_barrier(bio))
+  return -EOPNOTSUPP;
+#endif
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ gp = ioband_group_get(gp, bio);
+ prevent_burst_bios(gp);
+ if (should_pushback_bio(gp)) {
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  return DM_MAPIO_REQUEUE;
+ }
+
+ bio->bi_bdev = gp->c_dev->bdev;
+ bio->bi_sector -= ti->begin;
+retry:
+ if (!gp->c_blocked && room_for_bio(gp)) {
+  if (dp->g_can_submit(gp)) {
+   bws = &gp->c_stat[bio_data_dir(bio)];
+   bws->sectors += bio_sectors(bio);
```

```
+   bws->immediate++;
+   prepare_to_issue(gp, bio);
+   spin_unlock_irqrestore(&dp->g_lock, flags);
+   return DM_MAPIO_REMAPPED;
+  } else if (dp->g_issued < dp->g_io_throttle) {
+   dprintk(KERN_ERR "ioband_map: token expired "
+       "gp:%p bio:%p\n", gp, bio);
+   if (dp->g_restart_bios(dp))
+    goto retry;
+  }
+ }
+ hold_bio(gp, bio);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+
+ return DM_MAPIO_SUBMITTED;
+}
+
+/*
+ * Select the best group to resubmit its BIOs.
+ */
+static inline struct ioband_group *choose_best_group(struct ioband_device *dp)
+{
+ struct ioband_group *gp;
+ struct ioband_group *best = NULL;
+ int highest = 0;
+ int pri;
+
+ /* Todo: The algorithm should be optimized.
+  *       It would be better to use rbtree.
+  */
+ list_for_each_entry(gp, &dp->g_groups, c_list) {
+  if (!gp->c_blocked || !room_for_bio(gp))
+   continue;
+  pri = dp->g_can_submit(gp);
+  if (pri > highest) {
+   highest = pri;
+   best = gp;
+  }
+ }
+
+ return best;
+}
+
+/*
+ * This function is called right after it becomes able to resubmit BIOs.
+ * It selects the best BIOs and passes them to the underlying layer.
+ */
+#if LINUX_VERSION_CODE < KERNEL_VERSION(2,6,24)
```

```
+static void ioband_conduct(void *p)
+{
+ struct ioband_device *dp = (struct ioband_device *)p;
+#else
+static void ioband_conduct(struct work_struct *work)
+{
+ struct ioband_device *dp =
+  container_of(work, struct ioband_device, g_conductor);
+#endif
+ struct ioband_group *gp = NULL;
+ struct bio *bio;
+ unsigned long flags;
+ struct bio_list issue_list, pushback_list;
+
+ bio_list_init(&issue_list);
+ bio_list_init(&pushback_list);
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+retry:
+ while (dp->g_blocked && (gp = choose_best_group(dp)))
+  release_bios(gp, &issue_list, &pushback_list);
+
+ if (dp->g_blocked && dp->g_issued < dp->g_io_throttle) {
+  dprintk(KERN_ERR "ioband_conduct: token expired gp:%p\n", gp);
+  if (dp->g_restart_bios(dp))
+   goto retry;
+ }
+
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+
+ while ((bio = bio_list_pop(&issue_list)))
+  generic_make_request(bio);
+ while ((bio = bio_list_pop(&pushback_list)))
+#if LINUX_VERSION_CODE < KERNEL_VERSION(2,6,24)
+  bio_endio(bio, bio->bi_size, -EIO);
+#else
+  bio_endio(bio, -EIO);
+#endif
+}
+
+static int ioband_end_io(struct dm_target *ti, struct bio *bio,
+      int error, union map_info *map_context)
+{
+ struct ioband_group *gp = ti->private;
+ struct ioband_device *dp = gp->c_banddev;
+ unsigned long flags;
+ int r = error;
+
```

```
+ /*
+  *  XXX: A new error code for device mapper devices should be used
+  *       rather than EIO.
+  */
+ if (error == -EIO && should_pushback_bio(gp)) {
+  /* This ioband device is suspending */
+  r = DM_ENDIO_REQUEUE;
+ }
+ /* Todo: The algorithm should be optimized to eliminate the spinlock. */
+ spin_lock_irqsave(&dp->g_lock, flags);
+ if (dp->g_issued <= dp->g_io_throttle)
+  dp->g_plug_bio = 0;
+ dp->g_issued--;
+
+ /*
+  * Todo: It would be better to introduce high/low water marks here
+  *   not to kick the workqueues so often.
+  */
+ if (dp->g_blocked && !dp->g_plug_bio)
+  queue_work(dp->g_ioband_wq, &dp->g_conductor);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+ return r;
+}
+
+static void ioband_presuspend(struct dm_target *ti)
+{
+ struct ioband_group *gp = ti->private;
+ ioband_group_stop_all(gp, 1);
+}
+
+static void ioband_resume(struct dm_target *ti)
+{
+ struct ioband_group *gp = ti->private;
+ ioband_group_resume_all(gp);
+}
+
+
+static void ioband_group_status(struct ioband_group *gp, int *szp, char *result,
+    unsigned int maxlen)
+{
+ struct ioband_group_stat *bws;
+ unsigned long reqs;
+ int i, sz = *szp; /* used in DMEMIT() */
+
+ DMEMIT(" %d", gp->c_id);
+ for (i = 0; i < 2; i++) {
+  bws = &gp->c_stat[i];
+  reqs = bws->immediate + bws->deferred;
```

```
+  DMEMIT(" %lu %lu %lu",
+    bws->immediate + bws->deferred, bws->deferred,
+    bws->sectors);
+ }
+ *szp = sz;
+}
+
+static int ioband_status(struct dm_target *ti, status_type_t type,
+    char *result, unsigned int maxlen)
+{
+ struct ioband_group *gp = ti->private, *p;
+ struct ioband_device *dp = gp->c_banddev;
+ int sz = 0; /* used in DMEMIT() */
+ unsigned long flags;
+
+ switch (type) {
+ case STATUSTYPE_INFO:
+  spin_lock_irqsave(&dp->g_lock, flags);
+  DMEMIT("%d", dp->g_devgroup);
+  ioband_group_status(gp, &sz, result, maxlen);
+  list_for_each_entry(p, &gp->c_group_list, c_group_list)
+    ioband_group_status(p, &sz, result, maxlen);
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  break;
+
+ case STATUSTYPE_TABLE:
+  spin_lock_irqsave(&dp->g_lock, flags);
+  DMEMIT("%s %d %d %d", gp->c_dev->name, dp->g_devgroup,
+      dp->g_io_throttle, dp->g_io_limit);
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  break;
+ }
+
+ return 0;
+}
+
+static int ioband_group_type_select(struct ioband_group *gp, char *name)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct group_type *t;
+ unsigned long flags;
+
+ for (t = dm_ioband_group_type; (t->t_name); t++) {
+  if (!strcmp(name, t->t_name))
+    break;
+ }
+ if (!t->t_name) {
+  DMWARN("ioband type select: %s isn't supported.", name);
```

```
+  return -EINVAL;
+ }
+ spin_lock_irqsave(&dp->g_lock, flags);
+ if (!list_empty(&gp->c_group_list)) {
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  return -EBUSY;
+ }
+ gp->c_getid = t->t_getid;
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+
+ return 0;
+}
+
+static inline int split_string(char *s, char **v)
+{
+ int id = IOBAND_ID_ANY;
+ char *p, *q;
+
+ p = strsep(&s, "=:,");
+ q = strsep(&s, "=:,");
+ if (!q) {
+  *v = p;
+ } else {
+  id = simple_strtol(p, NULL, 0);
+  *v = q;
+ }
+ return id;
+}
+
+static int ioband_set_param(struct ioband_group *gp, char *cmd, char *value)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ char *val_str;
+ int id;
+ unsigned long flags;
+ int r;
+
+ id = split_string(value, &val_str);
+
+ spin_lock_irqsave(&dp->g_lock, flags);
+ if (id != IOBAND_ID_ANY) {
+  gp = ioband_group_find(gp, id);
+  if (!gp) {
+   spin_unlock_irqrestore(&dp->g_lock, flags);
+   DMWARN("ioband_set_param: id=%d not found.", id);
+   return -EINVAL;
+  }
+ }
```

```
+ r = dp->g_set_param(gp, cmd, val_str);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+ return r;
+}
+
+static int ioband_group_attach(struct ioband_group *gp, int id)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct ioband_group *sub_gp;
+ int r;
+
+ if (id < 0) {
+  DMWARN("ioband_group_attach: invalid id:%d", id);
+  return -EINVAL;
+ }
+ sub_gp = kzalloc(sizeof(struct ioband_group), GFP_KERNEL);
+ if (!sub_gp)
+  return -ENOMEM;
+
+ r = ioband_group_init(sub_gp, gp, dp, id);
+ if (r < 0) {
+  kfree(sub_gp);
+  return r;
+ }
+ return 0;
+}
+
+static int ioband_group_detach(struct ioband_group *gp, int id)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct ioband_group *sub_gp;
+ unsigned long flags;
+
+ if (id <= 0) {
+  DMWARN("ioband_group_detach: invalid id:%d", id);
+  return -EINVAL;
+ }
+ spin_lock_irqsave(&dp->g_lock, flags);
+ sub_gp = ioband_group_find(gp, id);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
+ if (!sub_gp) {
+  DMWARN("ioband_group_detach: invalid id:%d", id);
+  return -EINVAL;
+ }
+ ioband_group_stop(sub_gp);
+ spin_lock_irqsave(&dp->g_lock, flags);
+ ioband_group_release(sub_gp);
+ spin_unlock_irqrestore(&dp->g_lock, flags);
```

```
+ return 0;
+}
+
+/*
+ * Message parameters:
+ * "policy"      <name>
+ *       ex)
+ *  "policy" "weight"
+ * "type"        "none"|"pid"|"pgrp"|"node"|"cpuset"|"cgroup"|"user"|"gid"
+ *  "io_throttle" <value>
+ *  "io_limit"    <value>
+ * "attach"      <group id>
+ * "detach"      <group id>
+ * "any-command" <group id>:<value>
+ *       ex)
+ *  "weight" 0:<value>
+ *  "token"  24:<value>
+ */
+static int ioband_message(struct dm_target *ti, unsigned int argc, char **argv)
+{
+ struct ioband_group *gp = ti->private, *p;
+ struct ioband_device *dp = gp->c_banddev;
+ int val = 0;
+ int r = 0;
+ unsigned long flags;
+
+ if (argc == 1 && !strcmp(argv[0], "reset")) {
+  spin_lock_irqsave(&dp->g_lock, flags);
+  memset(gp->c_stat, 0, sizeof(gp->c_stat));
+  list_for_each_entry(p, &gp->c_group_list, c_group_list)
+   memset(p->c_stat, 0, sizeof(p->c_stat));
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  return 0;
+ }
+
+ if (argc != 2) {
+  DMWARN("Unrecognised band message received.");
+  return -EINVAL;
+ }
+ if (!strcmp(argv[0], "debug")) {
+  ioband_debug = simple_strtol(argv[1], NULL, 0);
+  if (ioband_debug < 0) ioband_debug = 0;
+  return 0;
+ } else if (!strcmp(argv[0], "io_throttle")) {
+  val = simple_strtol(argv[1], NULL, 0);
+  spin_lock_irqsave(&dp->g_lock, flags);
+  if (val <= 0 || val >= dp->g_io_limit) {
+   spin_unlock_irqrestore(&dp->g_lock, flags);
```

```
+   return -EINVAL;
+  }
+  dp->g_io_throttle = val;
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  ioband_set_param(gp, argv[0], argv[1]);
+  return 0;
+ } else if (!strcmp(argv[0], "io_limit")) {
+  val = simple_strtol(argv[1], NULL, 0);
+  spin_lock_irqsave(&dp->g_lock, flags);
+  if (val <= dp->g_io_throttle) {
+   spin_unlock_irqrestore(&dp->g_lock, flags);
+   return -EINVAL;
+  }
+  dp->g_io_limit = val;
+  spin_unlock_irqrestore(&dp->g_lock, flags);
+  ioband_set_param(gp, argv[0], argv[1]);
+  return 0;
+ } else if (!strcmp(argv[0], "type")) {
+  return ioband_group_type_select(gp, argv[1]);
+ } else if (!strcmp(argv[0], "attach")) {
+  int id = simple_strtol(argv[1], NULL, 0);
+  return ioband_group_attach(gp, id);
+ } else if (!strcmp(argv[0], "detach")) {
+  int id = simple_strtol(argv[1], NULL, 0);
+  return ioband_group_detach(gp, id);
+ } else {
+  /* message anycommand <group-id>:<value> */
+  r = ioband_set_param(gp, argv[0], argv[1]);
+  if (r < 0)
+   DMWARN("Unrecognised band message received.");
+  return r;
+ }
+ return 0;
+}
+
+static struct target_type ioband_target = {
+ .name      = "ioband",
+ .module      = THIS_MODULE,
+ .version     = {0, 0, 3},
+ .ctr       = ioband_ctr,
+ .dtr       = ioband_dtr,
+ .map       = ioband_map,
+ .end_io      = ioband_end_io,
+ .presuspend  = ioband_presuspend,
+ .resume      = ioband_resume,
+ .status      = ioband_status,
+ .message     = ioband_message,
+};
```

```
+
+static int __init dm_ioband_init(void)
+{
+ int r;
+
+ r = dm_register_target(&ioband_target);
+ if (r < 0) {
+  DMERR("register failed %d", r);
+  return r;
+ }
+ return r;
+}
+
+static void __exit dm_ioband_exit(void)
+{
+ int r;
+
+ r = dm_unregister_target(&ioband_target);
+ if (r < 0)
+  DMERR("unregister failed %d", r);
+}
+
+module_init(dm_ioband_init);
+module_exit(dm_ioband_exit);
+
+MODULE_DESCRIPTION(DM_NAME " I/O bandwidth control");
+MODULE_AUTHOR("Hirokazu Takahashi <taka@valinux.co.jp>, "
+       "Ryo Tsuruta <ryov@valinux.co.jp");
+MODULE_LICENSE("GPL");
diff -uprN linux-2.6.24.orig/drivers/md/dm-ioband-policy.c
linux-2.6.24/drivers/md/dm-ioband-policy.c
--- linux-2.6.24.orig/drivers/md/dm-ioband-policy.c 1970-01-01 09:00:00.000000000 +0900
+++ linux-2.6.24/drivers/md/dm-ioband-policy.c 2008-02-05 19:09:41.000000000 +0900
@@ -0,0 +1,202 @@
+/*
+ * Copyright (C) 2008 VA Linux Systems Japan K.K.
+ *
+ *  I/O bandwidth control
+ *
+ * This file is released under the GPL.
+ */
+#include <linux/bio.h>
+#include <linux/workqueue.h>
+#include "dm.h"
+#include "dm-bio-list.h"
+#include "dm-ioband.h"
+
+/*
```

```
+ * The following functions determine when and which BIOs should
+ * be submitted to control the I/O flow.
+ * It is possible to add a new BIO scheduling policy with it.
+ */
+
+
+/*
+ * Functions for weight balancing policy.
+ */
+#define DEFAULT_WEIGHT 100
+#define DEFAULT_TOKENBASE 2048
+#define IOBAND_IOPRIO_BASE 100
+
+static int proceed_global_epoch(struct ioband_device *dp)
+{
+ dp->g_epoch++;
+#if 0 /* this will also work correctly */
+ if (dp->g_blocked)
+  queue_work(dp->g_ioband_wq, &dp->g_conductor);
+ return 0;
+#endif
+ dprintk(KERN_ERR "proceed_epoch %d --> %d\n",
+      dp->g_epoch-1, dp->g_epoch);
+ return 1;
+}
+
+static inline int proceed_epoch(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ if (gp->c_my_epoch != dp->g_epoch) {
+  gp->c_my_epoch = dp->g_epoch;
+  return 1;
+ }
+ return 0;
+}
+
+static inline int iopriority(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ int iopri;
+
+ iopri = gp->c_token*IOBAND_IOPRIO_BASE/gp->c_token_init_value + 1;
+ if (gp->c_my_epoch != dp->g_epoch)
+  iopri += IOBAND_IOPRIO_BASE;
+ if (is_group_down(gp))
+  iopri += IOBAND_IOPRIO_BASE*2;
+
```

```
+ return iopri;
+}
+
+static int is_token_left(struct ioband_group *gp)
+{
+ if (gp->c_token > 0)
+  return iopriority(gp);
+
+ if (proceed_epoch(gp) || is_group_down(gp)) {
+  gp->c_token = gp->c_token_init_value;
+  dprintk(KERN_ERR "refill token: gp:%p token:%d\n",
+       gp, gp->c_token);
+  return iopriority(gp);
+ }
+ return 0;
+}
+
+static void prepare_token(struct ioband_group *gp, struct bio *bio)
+{
+ gp->c_token--;
+}
+
+static void set_weight(struct ioband_group *gp, int new)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ struct ioband_group *p;
+
+ dp->g_weight_total += (new - gp->c_weight);
+ gp->c_weight = new;
+
+ list_for_each_entry(p, &dp->g_groups, c_list) {
+  /* Fixme: it might overflow */
+  p->c_token = p->c_token_init_value =
+     dp->g_token_base*p->c_weight/dp->g_weight_total + 1;
+  p->c_limit =
+     dp->g_io_limit*p->c_weight/dp->g_weight_total + 1;
+ }
+}
+
+static int policy_weight_ctr(struct ioband_group *gp)
+{
+ struct ioband_device *dp = gp->c_banddev;
+
+ gp->c_my_epoch = dp->g_epoch;
+ gp->c_weight = 0;
+ set_weight(gp, DEFAULT_WEIGHT);
+ return 0;
+}
```

```
+
+static void policy_weight_dtr(struct ioband_group *gp)
+{
+ set_weight(gp, 0);
+}
+
+static int policy_weight_param(struct ioband_group *gp, char *cmd, char *value)
+{
+ struct ioband_device *dp = gp->c_banddev;
+ int val = simple_strtol(value, NULL, 0);
+ int r = 0;
+
+ if (!strcmp(cmd, "weight")) {
+  if (val > 0)
+   set_weight(gp, val);
+  else
+   r = -EINVAL;
+ } else if (!strcmp(cmd, "token")) {
+  if (val > 0) {
+   dp->g_token_base = val;
+   set_weight(gp, gp->c_weight);
+  } else
+   r = -EINVAL;
+ } else if (!strcmp(cmd, "io_limit")) {
+  set_weight(gp, gp->c_weight);
+ } else {
+  r = -EINVAL;
+ }
+ return r;
+}
+
+static int is_queue_full(struct ioband_group *gp)
+{
+ return gp->c_blocked >= gp->c_limit;
+}
+
+
+/*
+ *  <Method>      <description>
+ * g_can_submit   : To determine whether a given group has the right to
+ *                  submit BIOs. The larger the return value the higher the
+ *                  priority to submit. Zero means it has no right.
+ * g_prepare_bio  : Called right before submitting each BIO.
+ * g_restart_bios : Called if this ioband device has some BIOs blocked but none
+ *                  of them can be submitted now. This method has to
+ *                  reinitialize the data to restart to submit BIOs and return
+ *                  0 or 1.
+ *                  The return value 0 means that it has become able to submit
```

```
+ *               them now so that this ioband device will continue its work.
+ *               The return value 1 means that it is still unable to submit
+ *               them so that this device will stop its work. And this
+ *               policy module has to reactivate the device when it gets
+ *               to be able to submit BIOs.
+ * g_hold_bio    : To hold a given BIO until it is submitted.
+ *               The default function is used when this method is undefined.
+ * g_pop_bio     : To select and get the best BIO to submit.
+ * g_group_ctr   : To initalize the policy own members of struct ioband_group.
+ * g_group_dtr   : Called when struct ioband_group is removed.
+ * g_set_param   : To update the policy own date.
+ *               The parameters can be passed through "dmsetup message"
+ *               command.
+ * g_should_block : Called every time this ioband device receive a BIO.
+ *               Return 1 if a given group can't receive any more BIOs,
+ *               otherwise return 0.
+ */
+static void policy_weight_init(struct ioband_device *dp)
+{
+ dp->g_can_submit = is_token_left;
+ dp->g_prepare_bio = prepare_token;
+ dp->g_restart_bios = proceed_global_epoch;
+ dp->g_group_ctr = policy_weight_ctr;
+ dp->g_group_dtr = policy_weight_dtr;
+ dp->g_set_param = policy_weight_param;
+ dp->g_should_block = is_queue_full;
+
+ dp->g_token_base = DEFAULT_TOKENBASE;
+ dp->g_epoch = 0;
+ dp->g_weight_total = 0;
+}
+/* weight balancing policy. --- End --- */
+
+
+static void policy_default_init(struct ioband_device *dp) /*XXX*/
+{
+ policy_weight_init(dp); /* temp */
+}
+
+struct policy_type dm_ioband_policy_type[] = {
+ {"default", policy_default_init},
+ {"weight", policy_weight_init},
+ {NULL,     policy_default_init}
+};
diff -uprN linux-2.6.24.orig/drivers/md/dm-ioband-type.c linux-2.6.24/drivers/md/dm-ioband-type.c
--- linux-2.6.24.orig/drivers/md/dm-ioband-type.c 1970-01-01 09:00:00.000000000 +0900
+++ linux-2.6.24/drivers/md/dm-ioband-type.c 2008-02-05 19:09:41.000000000 +0900
@@ -0,0 +1,80 @@
```

```
+/*
+ * Copyright (C) 2008 VA Linux Systems Japan K.K.
+ *
+ *  I/O bandwidth control
+ *
+ * This file is released under the GPL.
+ */
+#include <linux/bio.h>
+#include "dm.h"
+#include "dm-bio-list.h"
+#include "dm-ioband.h"
+
+/*
+ * Any I/O bandwidth can be divided into several bandwidth groups, each of which
+ * has its own unique ID. The following functions are called to determine
+ * which group a given BIO belongs to and return the ID of the group.
+ */
+
+/* ToDo: unsigned long value would be better for group ID */
+
+static int ioband_process_id(struct bio *bio)
+{
+ /*
+  * This function will work for KVM and Xen.
+  */
+ return (int)current->tgid;
+}
+
+static int ioband_process_group(struct bio *bio)
+{
+#if LINUX_VERSION_CODE < KERNEL_VERSION(2,6,24)
+ return (int)process_group(current);
+#else
+ return (int)task_pgrp_nr(current);
+#endif
+}
+
+static int ioband_uid(struct bio *bio)
+{
+ return (int)current->uid;
+}
+
+static int ioband_gid(struct bio *bio)
+{
+ return (int)current->gid;
+}
+
+static int ioband_cpuset(struct bio *bio)
```

```
+{
+ return 0; /* not implemented yet */
+}
+
+static int ioband_node(struct bio *bio)
+{
+ return 0; /* not implemented yet */
+}
+
+static int ioband_cgroup(struct bio *bio)
+{
+  /*
+   * This function should return the ID of the cgroup which issued "bio".
+   * The ID of the cgroup which the current process belongs to won't be
+   * suitable ID for this purpose, since some BIOs will be handled by kernel
+   * threads like aio or pdflush on behalf of the process requesting the BIOs.
+   */
+ return 0; /* not implemented yet */
+}
+
+struct group_type dm_ioband_group_type[] = {
+ {"none",   NULL},
+ {"pgrp",   ioband_process_group},
+ {"pid",    ioband_process_id},
+ {"node",   ioband_node},
+ {"cpuset", ioband_cpuset},
+ {"cgroup", ioband_cgroup},
+ {"user",   ioband_uid},
+ {"uid",    ioband_uid},
+ {"gid",    ioband_gid},
+ {NULL,     NULL}
+};
diff -uprN linux-2.6.24.orig/drivers/md/dm-ioband.h linux-2.6.24/drivers/md/dm-ioband.h
--- linux-2.6.24.orig/drivers/md/dm-ioband.h 1970-01-01 09:00:00.000000000 +0900
+++ linux-2.6.24/drivers/md/dm-ioband.h 2008-02-05 19:09:41.000000000 +0900
@@ -0,0 +1,129 @@
+/*
+ * Copyright (C) 2008 VA Linux Systems Japan K.K.
+ *
+ *  I/O bandwidth control
+ *
+ * This file is released under the GPL.
+ */
+
+#include <linux/version.h>
+#include <linux/wait.h>
+
+#define DEFAULT_IO_THROTTLE 4
```

```
+#define DEFAULT_IO_LIMIT 128
+#define IOBAND_NAME_MAX 31
+#define IOBAND_ID_ANY (-1)
+
+struct ioband_group;
+
+struct ioband_device {
+ struct list_head g_groups;
+ struct work_struct g_conductor;
+ struct workqueue_struct *g_ioband_wq;
+ int g_io_throttle;
+ int g_io_limit;
+ int g_plug_bio;
+ int g_issued;
+ int g_blocked;
+ spinlock_t g_lock;
+ wait_queue_head_t g_waitq;
+
+ int g_devgroup;
+ int g_ref;  /* just for debugging */
+ struct list_head g_list;
+ int g_flags;
+ char g_name[IOBAND_NAME_MAX + 1]; /* rfu */
+
+ /* policy dependent */
+ int (*g_can_submit)(struct ioband_group *);
+ void (*g_prepare_bio)(struct ioband_group *, struct bio *);
+ int (*g_restart_bios)(struct ioband_device *);
+ void (*g_hold_bio)(struct ioband_group *, struct bio *);
+ struct bio * (*g_pop_bio)(struct ioband_group *);
+ int (*g_group_ctr)(struct ioband_group *);
+ void (*g_group_dtr)(struct ioband_group *);
+ int (*g_set_param)(struct ioband_group *, char *cmd, char *value);
+ int (*g_should_block)(struct ioband_group *);
+
+ /* members for weight balancing policy */
+ int g_epoch;
+ int g_weight_total;
+ int g_token_base;
+
+};
+
+struct ioband_group_stat {
+ unsigned long sectors;
+ unsigned long immediate;
+ unsigned long deferred;
+};
+
```

```
+struct ioband_group {
+ struct list_head c_list;
+ struct ioband_device *c_banddev;
+ struct dm_dev *c_dev;
+ struct dm_target *c_target;
+ struct bio_list c_blocked_bios;
+ struct list_head c_group_list;
+ int c_id; /* should be unsigned long or unsigned long long */
+ char c_name[IOBAND_NAME_MAX + 1]; /* rfu */
+ int c_blocked;
+ wait_queue_head_t c_waitq;
+ int c_flags;
+ struct ioband_group_stat c_stat[2]; /* hold rd/wr status */
+
+ /* type dependent */
+ int (*c_getid)(struct bio *);
+
+ /* members for weight balancing policy */
+ int c_weight;
+ int c_my_epoch;
+ int c_token;
+ int c_token_init_value;
+ int c_limit;
+
+ /* rfu */
+ /* struct bio_list c_ordered_tag_bios; */
+};
+
+#define DEV_BIO_BLOCKED 1
+
+#define set_device_blocked(dp)  ((dp)->g_flags |= DEV_BIO_BLOCKED)
+#define clear_device_blocked(dp) ((dp)->g_flags &= ~DEV_BIO_BLOCKED)
+#define is_device_blocked(dp)  ((dp)->g_flags & DEV_BIO_BLOCKED)
+
+
+#define IOG_BIO_BLOCKED 1
+#define IOG_GOING_DOWN 2
+#define IOG_SUSPENDED 4
+
+#define set_group_blocked(gp)  ((gp)->c_flags |= IOG_BIO_BLOCKED)
+#define clear_group_blocked(gp)  ((gp)->c_flags &= ~IOG_BIO_BLOCKED)
+#define is_group_blocked(gp)  ((gp)->c_flags & IOG_BIO_BLOCKED)
+
+#define set_group_down(gp)  ((gp)->c_flags |= IOG_GOING_DOWN)
+#define clear_group_down(gp)  ((gp)->c_flags &= ~IOG_GOING_DOWN)
+#define is_group_down(gp)  ((gp)->c_flags & IOG_GOING_DOWN)
+
+#define set_group_suspended(gp)  ((gp)->c_flags |= IOG_SUSPENDED)
```

```
+#define clear_group_suspended(gp) ((gp)->c_flags &= ~IOG_SUSPENDED)
+#define is_group_suspended(gp)  ((gp)->c_flags & IOG_SUSPENDED)
+
+struct policy_type {
+ const char *p_name;
+ void   (*p_policy_init)(struct ioband_device *);
+};
+
+extern struct policy_type dm_ioband_policy_type[];
+
+struct group_type {
+ const char *t_name;
+ int   (*t_getid)(struct bio *);
+};
+
+extern struct group_type dm_ioband_group_type[];
+
+/* Just for debugging */
+extern int ioband_debug;
+#define dprintk(format, a...) \
+ if (ioband_debug > 0) ioband_debug--, printk(format, ##a)
```

_____
Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers

---

Subject: [PATCH 2/2] dm-ioband v0.0.3: The I/O bandwidth controller: Document
Posted by Ryo Tsuruta on Tue, 05 Feb 2008 10:20:01 GMT
View Forum Message <> Reply to Message

Here is the document of dm-ioband.

Based on 2.6.24
Signed-off-by: Ryo Tsuruta <ryov@valinux.co.jp>
Signed-off-by: Hirokazu Takahashi <taka@valinux.co.jp>

```
diff -uprN linux-2.6.24.orig/Documentation/device-mapper/ioband.txt
linux-2.6.24/Documentation/device-mapper/ioband.txt
--- linux-2.6.24.orig/Documentation/device-mapper/ioband.txt 1970-01-01 09:00:00.000000000
+0900
+++ linux-2.6.24/Documentation/device-mapper/ioband.txt 2008-02-05 19:09:41.000000000
+0900
@@ -0,0 +1,728 @@
+=====================
+Document for dm-ioband
+=====================
+
```

+Contents:
+ What's dm-ioband all about?
+ Differences from the CFQ I/O scheduler
+ How dm-ioband works
+ Setup and Installation
+ Getting started
+ Command Reference
+ Examples
+ TODO
+
+
+What's dm-ioband all about?
+==========================
+dm-ioband is an I/O bandwidth controller implemented as a device-mapper driver.
+Several jobs using the same physical device have to share the bandwidth of
+the device. dm-ioband gives bandwidth to each job according to its weight,
+which each job can set its own value to.
+
+At this time, a job is a group of processes with the same pid or pgrp or uid.
+There is also a plan to make it support cgroup. A job can also be a virtual
+machine such as KVM or Xen.
+
+ +------+ +------+ +------+   +------+ +------+ +------+
+ |cgroup| |cgroup| | the | | pid | | pid | | the | jobs
+ | A | | B | |others| | X | | Y | |others|
+ +--|---+ +--|---+ +--|---+   +--|---+ +--|---+ +--|---+
+ +--V----+---V---+----V---+   +--V----+---V---+----V---+
+ | group | group | default| | group | group | default| ioband groups
+ | | | group | | | | group |
+ +-------+-------+--------+   +-------+-------+--------+
+ | ioband1 | | ioband2 | ioband devices
+ +-----------|-----------+   +-----------|-----------+
+ +-----------V-------------+-------------V-----------+
+ | | |
+ | sdb1 | sdb2 | physical devices
+ +-------------------------+-------------------------+
+
+
+Differences from the CFQ I/O scheduler
+======================================
+
+Dm-ioband is flexible to configure the bandwidth settings.
+
+Dm-ioband can work with any type of I/O scheduler such as the NOOP scheduler,
+which is often chosen for high-end storages, since it is implemented outside
+the I/O scheduling layer. It allows both of partition based bandwidth control
+and job --- a group of processes --- based control. In addition, it can
+set different configuration on each physical device to control its bandwidth.

+
+Meanwhile the current implementation of the CFQ scheduler has seven IO priority
+levels and all jobs whose processes have the same IO priority share the
+bandwidth assigned to this level between them. And IO priority is an attribute
+of a process so that it equally effects to all block devices.
+
+
+How dm-ioband works.
+====================
+Every ioband device has one ioband group, which by default is called the
+default group.
+
+Ioband devices can also have extra ioband groups in them. Each ioband group
+has a job to support and a weight. Proportional to the weight, dm-ioband gives
+tokens to the group.
+
+A group passes on I/O requests that its job issues to the underlying
+layer so long as it has tokens left, while requests are blocked
+if there aren't any tokens left in the group. One token is consumed each
+time the group passes on a request. dm-ioband will refill groups with tokens
+once all of groups that have requests on a given physical device use up their
+tokens.
+
+With this approach, a job running on an ioband group with large weight is
+guaranteed to be able to issue a large number of I/O requests.
+
+
+Setup and Installation
+======================
+
+Build a kernel with these options enabled:
+
+ CONFIG_MD
+ CONFIG_BLK_DEV_DM
+ CONFIG_DM_IOBAND
+
+If compiled as module, use modprobe to load dm-ioband.
+
+ # make modules
+ # make modules_install
+ # depmod -a
+ # modprobe dm-ioband
+
+"dmsetup targets" command shows all available device-mapper targets.
+"ioband" is displayed if dm-ioband has been loaded.
+
+ # dmsetup targets
+ ioband   v0.0.3

+
+
+Getting started
+===============
+The following is a brief description how to control the I/O bandwidth of
+disks. In this description, we'll take one disk with two partitions as an
+example target.
+
+
+Create and map ioband devices
+-----------------------------
+Create two ioband devices "ioband1" and "ioband2" and map them to "/dev/sda1"
+and "/dev/sda2" respectively.
+
+ # echo "0 $(blockdev --getsize /dev/sda1) ioband /dev/sda1 1" | \
+   dmsetup create ioband1
+ # echo "0 $(blockdev --getsize /dev/sda2) ioband /dev/sda2 1" | \
+   dmsetup create ioband2
+
+If the commands are successful then the device files "/dev/mapper/ioband1"
+and "/dev/mapper/ioband2" will have been created.
+
+
+Bandwidth control
+-----------------
+In this example, weights of 40 and 10 will be assigned to "ioband1" and
+"ioband2" respectively. This is done using the following commands:
+
+ # dmsetup message ioband1 0 weight 40
+ # dmsetup message ioband2 0 weight 10
+
+After these commands, "ioband1" can use 80% --- 40/(40+10)*100 --- of the
+bandwidth of the physical disk "/dev/sda" while "ioband2" can use 20%.
+
+
+Additional bandwidth control
+----------------------------
+In this example two extra ioband groups are created on "ioband1".
+The first group consists of all the processes with user-id 1000 and the
+second group consists of all the processes with user-id 2000. Their
+weights are 30 and 20 respectively.
+
+ # dmsetup message ioband1 0 type user
+ # dmsetup message ioband1 0 attach 1000
+ # dmsetup message ioband1 0 attach 2000
+ # dmsetup message ioband1 0 weight 1000:30
+ # dmsetup message ioband1 0 weight 2000:20
+

+Now the processes in the user-id 1000 group can use 30% ---
+30/(30+20+40+10)*100 --- of the bandwidth of the physical disk.
+
+ ioband device ioband group weight
+ ioband1 user id 1000 30
+ ioband1 user id 2000 20
+ ioband1 default group(the other users) 40
+ ioband2 default group 10
+
+
+Remove the ioband devices
+------------------------
+Remove the ioband devices when no longer used.
+
+ # dmsetup remove ioband1
+ # dmsetup remove ioband2
+
+
+Command Reference
+================
+
+
+Create an ioband device
+----------------------
+SYNOPSIS
+ dmsetup create IOBAND_DEVICE
+
+DESCRIPTION
+ Create an ioband device with the given name IOBAND_DEVICE. The following
+ arguments, which dmsetup command reads from standard input, are also
+ required.
+
+   Logical starting sector. This must be "0."
+   The number of sectors to use.
+   "ioband" as a target type.
+   The path name of the physical device.
+   Device group ID.
+   I/O throttling value (optional)
+   I/O limiting value (optional)
+
+ The same device group ID must be set among the ioband devices that share
+ the same bandwidth, which means they work on the same physical disk.
+ "The number of sectors to use" should be "the number of sectors the physical
+ device." I/O throttling value and I/O limiting value, which are described
+ later in this document, are optional.
+
+ If the command is successful, the device file
+ "/dev/device-mapper/IOBAND_DEVICE" will have been created.

+ An ioband group is also created and attached to IOBAND_DEVICE as the default
+ ioband group.
+
+EXAMPLE
+ Create an ioband device with the following parameters:
+   physical device = "/dev/sda1"
+   ioband device name = "ioband1"
+   device group ID = "1"
+   I/O throttling value = "10"
+   I/O limiting value = "200"
+
+   # echo "0 $(blockdev --getsize /dev/sda1) ioband /dev/sda1 1 10 200" | \
+   dmsetup create ioband1
+
+ Create two device groups (ID=1,2). The bandwidths of these device groups
+ will be individually controlled.
+
+   # echo "0 $(blockdev --getsize /dev/sda1) ioband /dev/sda1 1" | \
+   dmsetup create ioband1
+   # echo "0 $(blockdev --getsize /dev/sda2) ioband /dev/sda2 1" | \
+   dmsetup create ioband2
+   # echo "0 $(blockdev --getsize /dev/sdb3) ioband /dev/sdb3 2" | \
+   dmsetup create ioband3
+   # echo "0 $(blockdev --getsize /dev/sdb4) ioband /dev/sdb4 2" | \
+   dmsetup create ioband4
+
+
+Remove the ioband device
+------------------------
+SYNOPSIS
+ dmsetup remove IOBAND_DEVICE
+
+DESCRIPTION
+ Remove the specified ioband device IOBAND_DEVICE. All the band groups
+ attached to the ioband device are also removed automatically.
+
+EXAMPLE
+ Remove ioband device "ioband1."
+
+ # dmsetup remove ioband1
+
+
+Set an ioband group type
+-------------------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 type TYPE
+
+DESCRIPTION

+ Set the ioband group type of the specified ioband device IOBAND_DEVICE. TYPE
+ must be one of "user", "gid", "pid" or "pgrp." Once the type is set, new
+ ioband groups can be created on IOBAND_DEVICE.
+
+EXAMPLE
+ Set the ioband group type of ioband device "ioband1" to "user."
+
+ # dmsetup message ioband1 0 type user
+
+
+Create an ioband group
+----------------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 attach ID
+
+DESCRIPTION
+ Create an ioband group and attach it to IOBAND_DEVICE.
+ ID specifies user-id, group-id, process-id or process-group-id depending
+ the ioband group type of IOBAND_DEVICE.
+
+EXAMPLE
+ Create an ioband group which consists of all processes with user-id 1000 and
+ attach it to ioband device "ioband1."
+
+ # dmsetup message ioband1 0 type user
+ # dmsetup message ioband1 0 attach 1000
+
+
+Detach the ioband group
+-----------------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 detach ID
+
+DESCRIPTION
+ Detach the ioband group specified by ID from ioband device IOBAND_DEVICE.
+
+EXAMPLE
+ Detach the ioband group with ID "2000" from ioband device "ioband2."
+
+ # dmsetup message ioband2 0 detach 1000
+
+
+Set the weight of an ioband group
+---------------------------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 weight VAL
+ dmsetup message IOBAND_DEVICE 0 weight ID:VAL
+

```
+DESCRIPTION
+  Set the weight of the ioband group specified by ID. Set the weight of the
+  default ioband group of IOBAND_DEVICE if ID isn't specified.
+  The following example means that "ioband1" can use 80% --- 40/(40+10)*100
+   --- of the bandwidth of the physical disk while "ioband2" can use 20%.
+
+    # dmsetup message ioband1 0 weight 40
+    # dmsetup message ioband2 0 weight 10
+
+  The following lines have the same effect as the above:
+
+    # dmsetup message ioband1 0 weight 4
+    # dmsetup message ioband2 0 weight 1
+
+  VAL must be an integer larger than 0. The default value, which is assigned
+  to newly created ioband groups, is 100.
+
+EXAMPLE
+  Set the weight of the default ioband group of "ioband1" to 40.
+
+  # dmsetup message ioband1 0 weight 40
+
+  Set the weight of the ioband group of "ioband1" with ID "1000" to 10.
+
+  # dmsetup message ioband1 0 weight 1000:10
+
+
+Set the number of tokens
+------------------------
+SYNOPSIS
+  dmsetup message IOBAND_DEVICE 0 token VAL
+
+DESCRIPTION
+  Set the number of tokens to VAL. According to their weight, this number of
+  tokens will be distributed to all the ioband groups on the physical device
+  to which ioband device IOBAND_DEVICE belongs when they use up their tokens.
+
+  VAL must be an integer greater than 0. The default is 2048.
+
+EXAMPLE
+  Set the number of tokens of the physical device to which "ioband1" belongs
+  to 256.
+
+  # dmsetup message ioband1 0 token 256
+
+
+Set I/O throttling
+------------------
```

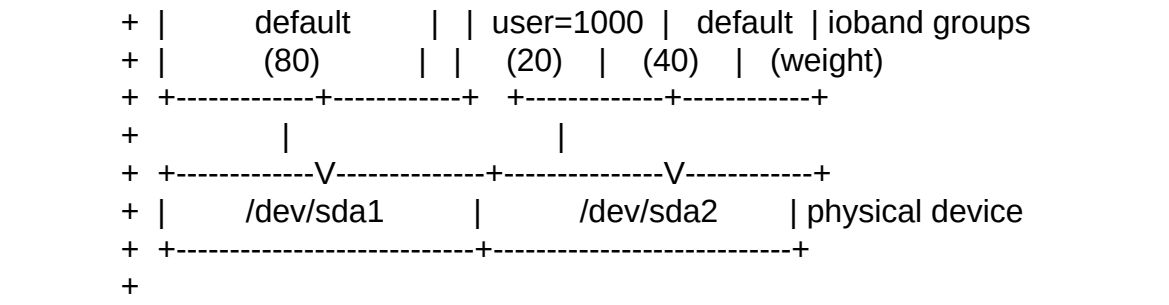
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 io_throttle VAL
+
+DESCRIPTION
+ Set the I/O throttling value of the physical disk to which ioband device
+ IOBAND_DEVICE belongs to VAL. Dm-ioband start to control the bandwidth
+ when the number of BIOs in progress on the physical disk exceeds this value.
+
+EXAMPLE
+ Set the I/O throttling value of "ioband1" to 16.
+
+ # dmsetup message ioband1 0 io_throttle 16
+
+
+Set I/O limiting
+----------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 io_limit VAL
+
+DESCRIPTION
+ Set the I/O limiting value of the physical disk to which ioband device
+ IOBAND_DEVICE belongs to VAL. Dm-ioband will block all I/O requests for
+ the physical device if the number of BIOs in progress on the physical disk
+ exceeds this value.
+
+EXAMPLE
+ Set the I/O limiting value of "ioband1" to 128.
+
+ # dmsetup message ioband1 0 io_limit 128
+
+
+Display settings
+----------------
+SYNOPSIS
+ dmsetup table --target ioband
+
+DESCRIPTION
+ Display the settings of all the ioband devices whose target type is "ioband."
+
+ The output format is as below:
+   ioband device name
+   starting sector of partition
+   partition size in sectors
+   target type
+   device number (major:minor)
+   device group ID
+   I/O throttle
+   I/O limit

+
+EXAMPLE
+ Display the setting of an ioband device configured such as:
+   device name = "ioband1"
+   starting sector of partition = "0"
+   partition size in sectors = "44371467"
+   target type = "ioband"
+   device number (major:minor) = "202:33"
+   device group ID = "128"
+   I/O throttle = "10"
+   I/O limit = "400"
+
+ # dmsetup table --target ioband
+ ioband1: 0 44371467 ioband 202:33 128 10 400
+
+Display Statistics
+------------------
+SYNOPSIS
+ dmsetup status --target ioband
+
+DESCRIPTION
+ Display the statistics of all the ioband devices whose target type is
+ "ioband."
+
+ The output format is as below. the first five columns shows:
+   ioband device name
+   logical start sector of the device (must be 0)
+   device size in sectors
+   target type (must be "ioband")
+   device group ID
+
+   The remaining columns show the statistics of each ioband group on the
+   band device. Each group uses seven columns for its statistics.
+   ioband group ID (-1 means default)
+   total read requests
+   delayed read requests
+   total read sectors
+   total write requests
+   delayed write requests
+   total write sectors
+
+EXAMPLE
+ The following output shows the statistics of two ioband devices. Ioband2 only
+ has the default ioband group and ioband1 has three (default, 1001, 1002)
+ ioband groups.
+
+ # dmsetup status
+ ioband2: 0 44371467 ioband 128 -1 143 90 424 122 78 352

+ ioband1: 0 44371467 ioband 128 -1 223 172 408 211 136 600 1001 166 107 \
+ 472 139 95 352 1002 211 146 520 210 147 504
+
+Reset status counter
+--------------------
+SYNOPSIS
+ dmsetup message IOBAND_DEVICE 0 reset
+
+DESCRIPTION
+ Reset the statistics of ioband device IOBAND_DEVICE.
+
+EXAMPLE
+ Reset the statistics of "ioband1."
+
+ # dmsetup message ioband1 0 reset
+
+
+Examples
+========
+
+
+Example #1: Bandwidth control on Partitions
+---
+This example describes how to control the bandwidth with disk partitions.
+The following diagram illustrates the configuration of this example.
+You may want to run a database on /dev/mapper/ioband1 and web applications
+on /dev/mapper/ioband2.
+
+   /mnt1   /mnt2   mount points
+ | |
+ +-------------V------------+   +-------------V------------+
+ | /dev/mapper/ioband1 | | /dev/mapper/ioband2 | ioband devices
+ +--------------------------+   +--------------------------+
+ | default group | | default group | ioband groups
+ | (80) | | (40) | (weight)
+ +-------------|------------+   +-------------|------------+
+ | |
+ +-------------V--------------+---------------V------------+
+ | /dev/sda1 | /dev/sda2 | physical devices
+ +----------------------------+----------------------------+
+
+To setup the above configuration, follow these steps:
+
+ 1) Create ioband devices with the same device group ID.
+
+   # echo "0 $(blockdev --getsize /dev/sda1) ioband /dev/sda1 1" | \
+   dmsetup create ioband1
+   # echo "0 $(blockdev --getsize /dev/sda2) ioband /dev/sda2 1" | \

+   dmsetup create ioband2
+
+ 2) Assign weights of 80 and 40 to the default ioband groups respectively.
+
+   # dmsetup message ioband1 0 weight 80
+   # dmsetup message ioband2 0 weight 40
+
+ 3) Create filesystems on the ioband devices and mount them.
+
+   # mkfs.ext3 /dev/mapper/ioband1
+   # mount /dev/mapper/ioband1 /mnt1
+
+   # mkfs.ext3 /dev/mapper/ioband2
+   # mount /dev/mapper/ioband2 /mnt2
+
+
+
+Example #2: Bandwidth control on Logical Volumes
+---
+This example is similar to the example #1 but it uses LVM logical volumes
+instead of disk partitions. This example shows how to configure ioband devices
+on two striped logical volumes.
+
+   /mnt1   /mnt2   mount points
+ | |
+ +------------V-----------+   +------------V-----------+
+ | /dev/mapper/ioband1 | | /dev/mapper/ioband2 | ioband devices
+ +--------------------------+   +--------------------------+
+ | default group | | default group | ioband groups
+ | (80) | | (40) | (weight)
+ +------------|-----------+   +------------|-----------+
+ | |
+ +------------V-----------+   +------------V-----------+
+ | /dev/mapper/lv0 | | /dev/mapper/lv1 | striped logical
+ | | | | volumes
+ +---+
+ | vg0 | volume group
+ +------------|-----------------------------|-----------+
+ | |
+ +------------V-----------+   +------------V-----------+
+ | /dev/sdb | | /dev/sdc | physical devices
+ +--------------------------+   +--------------------------+
+
+To setup the above configuration, follow these steps:
+
+ 1) Initialize the partitions for use by LVM.
+
+   # pvcreate /dev/sdb

```
+    # pvcreate /dev/sdc
+
+  2) Create a new volume group named "vg0" with /dev/sdb and /dev/sdc.
+
+    # vgcreate vg0 /dev/sdb /dev/sdc
+
+  3) Create two logical volumes in "vg0." The volumes have to be striped.
+
+    # lvcreate -n lv0 -i 2 -l 64 vg0 -L 1024M
+    # lvcreate -n lv1 -i 2 -l 64 vg0 -L 1024M
+
+The rest is the same as the example #1.
+
+  4) Create ioband devices corresponding to each logical volume.
+
+    # echo "0 $(blockdev --getsize /dev/mapper/lv0) ioband /dev/mapper/lv0 1"\
+       | dmsetup create ioband1
+    # echo "0 $(blockdev --getsize /dev/mapper/lv1) ioband /dev/mapper/lv1 1"\
+       | dmsetup create ioband2
+
+  5) Assign weights of 80 and 40 to the default ioband groups respectively.
+
+    # dmsetup message ioband1 0 weight 80
+    # dmsetup message ioband2 0 weight 40
+
+  6) Create filesystems on the ioband devices and mount them.
+
+    # mkfs.ext3 /dev/mapper/ioband1
+    # mount /dev/mapper/ioband1 /mnt1
+
+    # mkfs.ext3 /dev/mapper/ioband2
+    # mount /dev/mapper/ioband2 /mnt2
+
+
+
+Example #3: Bandwidth control on processes
+----------------------------------------
+This example describes how to control the bandwidth with groups of processes.
+You may also want to run an additional application on the same machine
+described in the example #1. This example shows how to add a new ioband group
+for this application.
+
+
+            /mnt1                    /mnt2          mount points
+              |                        |
+  +-------------V------------+   +-------------V------------+
+  |   /dev/mapper/ioband1    |   |   /dev/mapper/ioband2    | ioband devices
+  +--------------+-----------+   +--------------+-----------+
```



The following shows to set up a new ioband group on the machine that is already configured as the example #1. The application will have a weight of 20 and run with user-id 1000 on /dev/mapper/ioband2.

1) Set the type of ioband2 to "user."

```
# dmsetup message ioband2 0 type user.
```

2) Create a new ioband group on ioband2.

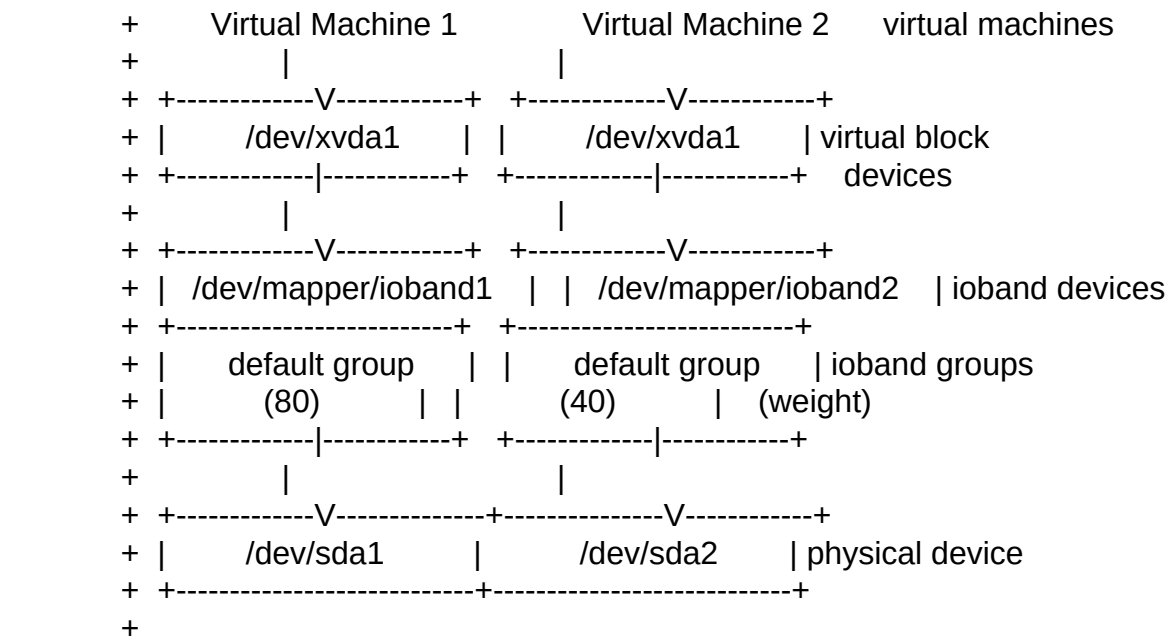
```
# dmsetup message ioband2 0 attach 1000
```

3) Assign weight of 10 to this newly created ioband group.

```
# dmsetup message ioband2 0 weight 1000:20
```

Example #4: Bandwidth control for Xen virtual block devices

This example describes how to control the bandwidth for Xen virtual block devices. The following diagram illustrates the configuration of this example.



+The followings shows how to map ioband device "ioband1" and "ioband2" to
+virtual block device "/dev/xvda1 on Virtual Machine 1" and "/dev/xvda1 on
+Virtual Machine 2" respectively on the machine configured as the example #1.
+Add the following lines to the configuration files that are referenced when
+creating "Virtual Machine 1" and "Virtual Machine 2."
+
+   For "Virtual Machine 1"
+   disk = ['phy:/dev/mapper/ioband1,xvda,w']
+
+   For "Virtual Machine 2"
+   disk = ['phy:/dev/mapper/ioband2,xvda,w']
+
+
+
+Example #5: Bandwidth control for Xen blktap devices
+---
+This example describes how to control the bandwidth for Xen virtual
+block devices when Xen blktap devices are used. The following diagram
+illustrates the configuration of this example.
+
+   Virtual Machine 1   Virtual Machine 2   virtual machines
+ | |
+ +------------V------------+   +------------V------------+
+ | /dev/xvda1 | | /dev/xvda1 | virtual block
+ +-------------|-----------+   +-------------|-----------+   devices
+ | |
+ +------------V-----------------------------V------------+
+ | /dev/mapper/ioband1 | ioband device
+ +---------------------------+---------------------------+
+ | default group | default group | ioband groups
+ | (80) | (40) | (weight)
+ +-------------|-------------+---------------|-----------+
+ | |
+ +-------------|-----------------------------|-----------+
+ | +----------V----------+   +----------V---------+ |
+ | | vm1.img | | vm2.img | | disk image files
+ | +---------------------+   +--------------------+ |
+ | /vmdisk | mount point
+ +---------------------------|---------------------------+
+ |
+ +---------------------------V---------------------------+
+ | /dev/sda1 | physical device
+ +---+
+
+To setup the above configuration, follow these steps:
+
+ 1) Create an ioband device.
+

```
+    # echo "0 $(blockdev --getsize /dev/sda1) ioband /dev/sda1 1" | \
+        dmsetup create ioband1
+
+  2) Add the following lines to the configuration files that are referenced
+     when creating "Virtual Machine 1" and "Virtual Machine 2."
+     Disk image files "/vmdisk/vm1.img" and "/vmdisk/vm2.img" will be used.
+
+    For "Virtual Machine 1"
+    disk = [ 'tap:aio:/vmdisk/vm1.img,xvda,w', ]
+
+    For "Virtual Machine 1"
+    disk = [ 'tap:aio:/vmdisk/vm2.img,xvda,w', ]
+
+  3) Run the virtual machines.
+
+    # xm create vm1
+    # xm create vm2
+
+  4) Find out the process IDs of the daemons which control the blktap devices.
+
+    # lsof /vmdisk/disk[12].img
+    COMMAND   PID USER   FD   TYPE DEVICE       SIZE  NODE NAME
+    tapdisk 15011 root   11u   REG  253,0 2147483648 48961 /vmdisk/vm1.img
+    tapdisk 15276 root   13u   REG  253,0 2147483648 48962 /vmdisk/vm2.img
+
+  5) Create new ioband groups of pid 15011 and pid 15276, which are process
+     IDs of the tapdisks, and assign weight of 80 and 40 to the groups
+     respectively.
+
+    # dmsetup message ioband1 0 type pid
+    # dmsetup message ioband1 0 attach 15011
+    # dmsetup message ioband1 0 weight 15011:80
+    # dmsetup message ioband1 0 attach 15276
+    # dmsetup message ioband1 0 weight 15276:40
+
+
+
+
+
+TODO
+====
+  - Cgroup support.
+  - Create a mechanism to track down which I/O is originally issued by which
+    process or cgroup.
+  - Control read and write requests separately.
+  - Support WRITE_BARRIER.
+  - Hierarchical ioband groups support.
+  - Optimization.
+  - More configuration tools. Or is the dmsetup command sufficient?
```

+ - Other policies to schedule BIOs. Or is the weight policy sufficient?
+ Is a new policy for LUNs, which may partially shared the bandwidth, needed?
+ - Other accounting policies to determine the bandwidths. Or is the number
+ of BIOs sufficient?

_____

Containers mailing list
Containers@lists.linux-foundation.org
https://lists.linux-foundation.org/mailman/listinfo/containers