AIX Version 7.2

Remote Direct Memory Access

AIX Version 7.2

Remote Direct Memory Access

หมายเหตุ

ก่อนที่คุณจะใช้ข้อมูลนี้และผลิตภัณฑ์ที่สนับสนุน โปรดอ่านข้อมูลใน “คำประกาศ” ในหน้า 19

เอดิชันนี้ใช้กับ AIX เวอร์ชัน 7.2 และกับรีลีสและโมดิฟิเคชันในลำดับต่อมาทั้งหมด จนกว่ามีการบ่งชี้เป็นอย่างอื่นในเอดิชันใหม่

สารบัญ

เกี่ยวกับเอกสารนี้

เอกสารนี้ให้ข้อมูลโดยละเอียดแก่โปรแกรมเมอร์ C ผู้มีประสบการณ์ เกี่ยวกับการโปรแกรมมิงโดยใช้ Open Fabrics Enterprise Distribution (OFED) verbs บน Internet Wide Area RDMA Protocol (iWARP) หรือ RDMA Network Interface Controller (RNIC) fabrics ในระบบปฏิบัติการ AIX®

เมื่อต้องการใช้เอกสารอย่างมีประสิทธิภาพ คุณควรทำความคุ้นเคยกับคำสั่ง การเรียกใช้งานระบบ รูทีนย่อย รูปแบบไฟล์ และไฟล์พิเศษ

การเน้น

ระเบียบการไฮไลต์ต่อไปนี้ถูกใช้ในเอกสารนี้:

Bold	ระบุคำสั่ง รูทีนย่อย คีย์เวิร์ด ไฟล์ โครงสร้าง ไดเร็กทอรี และรายการอื่นๆ ที่มีชื่อ ถูกกำหนดไว้แล้วโดยระบบ รวมทั้งระบุอ็อบเจ็กต์กราฟิก เช่น ปุ่ม เลเบล และไอคอนที่ผู้ใช้เลือก
ตัวเอียง	ระบุพารามิเตอร์ที่ชื่อแท้จริง หรือค่าที่ถูกกำหนด โดยผู้ใช้
`Monospace`	ระบุตัวอย่างของค่าข้อมูลที่ระบุเฉพาะ ตัวอย่างของข้อความที่ดูคล้ายกับที่คุณอาจมองเห็นจากที่แสดง ระบุตัวอย่างของส่วนของโค้ดโปรแกรมที่ดูคล้ายกับที่คุณอาจเขียนในฐานะเป็นโปรแกรมเมอร์ ระบุข้อความจากระบบ หรือข้อมูลที่คุณควรจะพิมพ์

การคำนึงถึงขนาดตัวพิมพ์ใน AIX

ทุกสิ่งในระบบปฏิบัติการ AIX เป็นแบบตรงตาม ตัวพิมพ์ ซึ่งหมายความว่ามีการแยกแยะความแตกต่างระหว่างตัวอักษรพิมพ์ใหญ่ และพิมพ์เล็ก ตัวอย่าง คุณสามารถใช้คำสั่ง **ls** เพื่อแสดงรายการไฟล์ ถ้าคุณพิมพ์ `LS` ระบบจะตอบกลับคำสั่งนั้นว่า `is not found` นอกจากนั้น **FILEA, FiLea** และ **filea** คือชื่อไฟล์สามชื่อที่แตกต่างกัน แม้ว่า ชื่อเหล่านั้นจะอยู่ในไดเร็กทอรีเดียวกัน เพื่อหลีกเลี่ยงสาเหตุของกการดำเนินการที่ไม่ต้องการดำเนินการ ตรวจสอบให้แน่ใจว่า คุณใช้กรณีที่ถูกต้อง

ISO 9000

ระบบรับรองคุณภาพที่ลงทะเบียน ISO 9000 ใช้ในการพัฒนาและการผลิตผลิตภัณฑ์นี้

Remote Direct Memory Access

โปรแกรมเมอร์ C ที่มีประสบการณ์สามารถค้นหาข้อมูลรายละเอียด เกี่ยวกับโปรแกรมมิงด้วย Remote Direct Memory Access (RDMA) verbs และ Open Fabrics Enterprise Distribution (OFED) verbs ในระบบปฏิบัติการ AIX

เมื่อต้องการใช้ข้อมูลอย่างมีประสิทธิภาพ คุณต้องทำความคุ้นเคยกับคำสั่ง การเรียกใช้งานระบบ รูทีนย่อย รูปแบบไฟล์ และไฟล์พิเศษ

Open Fabrics Enterprise Distribution (OFED)

เรียนรู้วิธีเริ่มต้นใช้งานด้วยโปรแกรมมิง Open Fabrics Enterprise Distribution (OFED) verbs ในระบบปฏิบัติการ AIX OFED verbs อนุญาตให้แอ็พพลิเคชันที่ต้องการทรูพุตสูง และเวลาแฝงต่ำใช้คุณลักษณะ Remote Direct Memory Access (RDMA)

แนวคิดสำหรับ OFED

เลเยอร์ verb สำหรับ Open Fabrics Enterprise Distribution (OFED) verbs เป็นค่าทั่วไปสำหรับ InfiniBand, RDMA over Converged Ethernet (RoCE), Internet Wide Area RDMA Protocol (iWARP) และ verbs ที่ สืบทอดจากสถาปัตยกรรม InfiniBand

ข้อกำหนดฮาร์ดแวร์

ระบบปฏิบัติการ AIX สนับสนุนอะแด็ปเตอร์ RDMA over Converged Ethernet (RoCE) ฮาร์ดแวร์ที่สนับสนุน RoCE RDMA ใน AIX ถูกเรียกอแดปเตอร์ PCIe2 10 GbE RoCE

ข้อกำหนดเกี่ยวกับซอฟต์แวร์

AIX OFED Verbs อิงตามโค้ด OFED 1.5 ของ OpenFabrics Alliance แอ็พพลิเคชันผู้ใช้ 32 บิต และ 64 บิต ของโค้ด OFED ได้รับการสนับสนุนบนระบบปฏิบัติการ AIX ไลบรารี ต่อไปนี้ถูกนำส่งพร้อมกับการติดตั้ง RDMA:

- Librdmacm
- Libibverbs

Verbs API

แอ็พพลิเคชัน AIX สามารถกำหนด verbs API ที่เป็น Open Fabrics Enterprise Distribution (OFED) verbs หรือ AIX InfiniBand (IB) verbs ที่ต้องสื่อสารกับปลายทางที่เจาะจง

ตัวอย่างต่อไปนี้ในโค้ดจำลองทดสอบผลลัพธ์ของคำสั่ง **rdma_resolve_addr** บนรีโมตแอดเดรสที่ต้องการเพื่อพิจารณา OFED verbs ที่สามารถ ใช้ได้

โปรแกรมส่งคืนค่าต่อไปนี้:

- `0`- ถ้าการสื่อสารกับปลายทาง สามารถจัดสร้างขึ้นได้โดยใช้ OFED verbs
- `error`- ถ้าการสื่อสารกับปลายทาง ไม่สามารถจัดสร้างขึ้นผ่านอุปกรณ์ที่สนับสนุน OFED แต่การสื่อสาร สามารถจัดสร้างขึ้นโดยใช้สถาปัตยกรรม InfiniBand

```
/*The following check_ofed_verbs_support routine does:
/*- Call rdma_create_event_channel to open a channel event     */
/*- Calls rdma_create_id() to get a cm_id                        */
/*- And then calls rdma_resolve_addr()             */
/*- Get the communication event                */
/*- Returns the event status:                  */
/* 0: OK                               */
/* error: NOK  output device may be not a RNIC device     */
/*- Calls rdma_destroy_id() to delete the cm_id created     */
/*- Call rdma_destroy_event_channel  to close a channel event  */

 int check_ofed_verbs_support (struct sockaddr *remoteaddr)
   {
    struct rdma_event channel *cm_channel;
    struct rdma_cm_id *cm_id;
    int ret=0;
    cm_channel = rdma_create_event_channel();
    if (!cm_channel)  {
      fprintf(stderr,"rdma_create_event_channel error\n");
      return -1;
      }
      ret = rdma_create_id(cm_channel, &cm_id, NULL, RDMA_PS_TCP);
      if (ret) {
        fprintf(stderr,"rdma_create_id: %d\n", ret);
        rdma_destroy_event_channel(cm_channel);
        return(ret);
      }
      ret = rdma_resolve_addr(cm_id, NULL, remoteaddr, RESOLVE_TIMEOUT_MS);
      if (ret) {
        fprintf(stderr,"rdma_resolve_addr: %d\n", ret);
        goto out;
      }
      ret = rdma_get_cm_event(cm_channel, &event);
      if (ret) {
        fprintf(stderr," rdma_get_cm_event() failed\n");
        goto out;
      }
      ret = event->status;
      rdma_ack_cm_event(event);
         out:
      rdma_destroy_id(cm_id);
      rdma_destroy_event_channel(cm_channel);
      return(ret);
   }
```

ไลบรารี Libibverbs

ไลบรารี **Libibverbs** เปิดใช้งานกระบวนการในพื้นที่ผู้ใช้เพื่อใช้ Remote Direct Memory Access (RDMA) verbs

ไลบรารี **Libibverbs** ถูกอธิบายในข้อมูลจำเพาะด้านสถาปัตยกรรม InfiniBand และข้อมูลจำเพาะ verbs ของโปรโตคอล RDMA

โหนดอุปกรณ์อักขระ `/dev/rdma/uverbsN` หลายโหนด ถูกใช้เพื่อสื่อสารระหว่างไลบรารี **Libibverbs** และเคอร์เนลเลเยอร์ `ib_uverbs` ทุกอะแด็ปเตอร์ RDMA network interface controller (RNIC) มีหนึ่งอุปกรณ์ที่ถูกรีจิสเตอร์ กับ Open Fabrics Enterprise Distribution (OFED) core เช่น อุปกรณ์ uverbs1 และ uverbs2 เมื่อต้องการรันบนอุปกรณ์ที่เหมาะสม ไลบรารีจะเขียนคำสั่งที่เกี่ยวข้อง กับ verb

ข้อมูลที่เกี่ยวข้อง:

InfiniBand

RDMA protocol verbs

ไลบรารี Librdmacm

ไลบรารี **librdmacm** จัดให้มีฟังก์ชัน communication manager (CM) และชุดทั่วไปของอินเตอร์เฟส Remote Direct Memory Access (RDMA) CM ที่รันบนเฟบริกต่างกัน เช่น InfiniBand (IB), RDMA over Converged Ethernet (RoCE) หรือ Internet Wide Area RDMA Protocol (iWARP)

โหนดอุปกรณ์ `/dev/rdma/rdma_cm` เดียวถูกใช้โดยพื้นที่ผู้ใช้เพื่อสื่อสารกับเคอร์เนล ไม่ว่าจะมี จำนวนอะแด็ปเตอร์ หรือพอร์ตเท่าใด แสดงอยู่

ไลบรารี **librdmacm** ถูกใช้โดยแอ็พพลิเคชัน ที่ต้องรันบนอุปกรณ์ RDMA ใดๆ

RDMA network interface controller (RNIC)

อะแด็ปเตอร์ I/O เครือข่ายหรือตัวควบคุมที่ฝังกับ Internet Wide Area RDMA Protocol (iWARP) และฟังก์ชัน Verbs

ตัวจัดการการสื่อสาร RDMA_CM

ตัวจัดการการสื่อสาร Remote Direct Memory Access (RDMA_CM) ใช้เพื่อตั้งค่าการเชื่อมต่อที่เรพลิเคตได้สำหรับการถ่ายโอนข้อมูล

ตัวจัดการการสื่อสารมีส่วนติดต่อ RDMA transport neutral สำหรับการสร้างการเชื่อมต่อ API อิงตามซ็อกเก็ต แต่ถูก ปรับใช้สำหรับซีแมนทิกส์ตามคู่คิว (QP) การสื่อสารจะ ผ่านอุปกรณ์ RDMA ที่เจาะจง และการถ่ายโอนข้อมูลจะเป็นแบบข้อความ

RDMA CM ใช้ไลบรารี **librdmacm** เพื่อระบุ การจัดการการสื่อสารเพื่อตั้งค่า และ teardown การเชื่อมต่อ ของ RDMA API ตัวจัดการการสื่อสารทำงานกับ verbs API โดยใช้ไลบรารี **libibverbs** สำหรับการถ่ายโอนข้อมูล

รีซอร์สที่จัดการโดยใช้ OFED verbs

แสดงรายการรีซอร์สที่จัดการโดยใช้ OFED verbs

Completion Queue (CQ):

คิว first-in-first-out (FIFO) ที่มี Completion queues (CQ) CQ สัมพันธ์กับคู่คิว ซึ่งใช้เพื่อรับ การแจ้งเตือนการทำสำเร็จ และเหตุการณ์

Completion Queue Entry (CQE):

รายการใน CQ ที่อธิบายข้อมูลเกี่ยวกับ Work request (WR) ที่เสร็จสมบูรณ์ เช่นสถานะ และขนาด

Event Channel:

ใช้เพื่อรายงานเหตุการณ์การสื่อสาร แต่ละแชนเนลเหตุการณ์ถูกแม็พ กับ descriptor ไฟล์ Descriptor ไฟล์ที่สัมพันธ์กันสามารถใช้ และ ดำเนินการเช่นเดียวกับ descriptor ไฟล์อื่น เพื่อเปลี่ยนแปลงลักษณะการทำงาน คุณสามารถทำให้ descriptor ไฟล์ดำเนินการหนึ่งในแอ็คชันต่อไปนี้:

- ไม่บล็อก descriptor ไฟล์
- โพล descriptor ไฟล์
- เลือก descriptor ไฟล์

Memory Region (MR):

ชุดของบัฟเฟอร์หน่วยความจำที่รีจิสเตอร์โดยมีสิทธิ์การเข้าถึง ในการใช้บัฟเฟอร์หน่วยความจำด้วยอะแด็ปเตอร์เครือข่าย ส่วนของหน่วยความจำ ต้องถูกรีจิสเตอร์

Protection Domain (PD):

เปิดให้ไคลเอ็นต์เชื่อมโยงกับหลายรีซอร์ส เช่น คู่ คิว และส่วนหน่วยความจำ ภายในโดเมน จากนั้นไคลเอ็นต์ให้สิทธิ์การเข้าถึงเพื่อส่งหรือรับข้อมูลภายในโดเมนที่มีการปกป้อง แก่โดเมนอื่นที่อยู่บน RDMA fabric

Queue Pair (QP):

Queue pairs (QPs) มีคิวการส่งและรับ คิว ส่งจะส่งข้อความขาออกที่ร้องขอการดำเนินการ RDMA คิวรับจะรับข้อความขาเข้า หรือข้อมูลระหว่างกลาง

Scatter or Gather Elements (SGE):

รายการไปยังพอยเตอร์ที่ชี้ไปยังบล็อกหน่วยความจำที่รีจิสเตอร์โลคัล แบบเต็ม หรือบางส่วน อิลิเมนต์เก็บแอดเดรสเริ่มต้นของบล็อก ขนาด และ lkey ที่มีสิทธิ์ที่เกี่ยวข้อง

Scatter or Gather Array:

อาร์เรย์ของอิลิเมนต์ scatter หรือ gather ที่มีอยู่ใน work request (WR) อาร์เรย์ทำงานตามโค้ดดำเนินการที่ รวบรวมข้อมูลจากหลายๆ บัฟเฟอร์ และส่งเป็นสตรีมเดียว หรือใช้สตรีมเดียว และแบ่งข้อมูลออกเป็นหลายบัฟเฟอร์

Work Queue (WQ):

Work queue ประกอบด้วย Send Queue หรือ Receive Queue Work queue ใช้เพื่อส่งหรือรับข้อความ

Work Queue Element (WQE):

Work Queue Element เป็นอิลิเมนต์ใน work queue

Work Request (WR):

Work Request คือการร้องขอที่โพสต์โดยผู้ใช้ไปยัง work queue

การดำเนินการด้านการสื่อสาร

แสดงรายการการดำเนินการด้านการสื่อสารที่มีสำหรับ อุปกรณ์ RDMA

ส่ง และส่งด้วยการดำเนินการระหว่างกลาง:

การดำเนินการส่งจะส่งข้อมูลไปยังคิวการรับของ Queue Pair (QP) รีโมต

เมื่อต้องการรับข้อมูล ผู้รับต้องโพสต์ข้อมูลไปยัง บัฟเฟอร์รับ ผู้ส่งไม่มีการควบคุมใดๆ ในข้อมูล ที่อยู่ในรีโมตโฮสต์

ค่า 4 ไบต์ระหว่างกลางจะถูกส่ง ไปกับบัฟเฟอร์ข้อมูล ค่าระหว่างกลางนี้ถูกแสดงต่อผู้ใช้ เพื่อเป็นส่วนหนึ่งของการแจ้งการได้รับ และไม่มีอยู่ใน บัฟเฟอร์ข้อมูล

การดำเนินการรับ:

การดำเนินการรับคือการดำเนินการที่สอดคล้องกับ การดำเนินการส่ง

โฮสต์ที่ทำงานรับได้รับแจ้งว่าได้รับบัฟเฟอร์ข้อมูลที่มี ค่าระหว่างกลางที่อินไลน์ แอ็พพลิเคชันการรับจะเก็บรักษา บัฟเฟอร์การรับ และข้อมูลการโพสต์

การดำเนินการ RDMA read:

การดำเนินการ RDMA read อ่านในส่วน หน่วยความจำจากรีโมตโฮสต์

คุณต้องระบุแอดเดรสเสมือนรีโมต และแอดเดรสหน่วยความจำ โลคัลที่ข้อมูลการอ่านถูกคัดลอก ก่อนคุณรันการดำเนินการ Remote Direct Memory Access (RDMA) รีโมตโฮสต์ต้องระบุ สิทธิ์ที่เหมาะสมเพื่อเข้าถึงหน่วยความจำ หลังจากสิทธิ์ ถูกตั้งค่า การดำเนินการ RDMA read จะรันโดยไม่มีการแจ้งเตือน ไปยังรีโมตโฮสต์

การดำเนินการ Atomic:

การดำเนินการ Atomic ไม่ได้รับการสนับสนุนโดยฮาร์ดแวร์ Remote Direct Memory Access (RDMA) ที่มีสำหรับระบบปฏิบัติการ AIX

RDMA write หรือ RDMA write ที่มีการดำเนินการระหว่างกลาง:

การดำเนินการ RDMA write คล้ายกับการดำเนินการ RDMA read แต่ข้อมูลจะถูกเขียนไปยังรีโมตโฮสต์

การดำเนินการ RDMA write จะรันโดยไม่มีการแจ้งเตือนไปยังรีโมต โฮสต์ RDMA write ที่มีการดำเนินการระหว่างกลางจะแจ้งรีโมตโฮสต์ ให้ทราบเกี่ยวกับค่าระหว่างกลาง

โหมดการขนส่ง

โหมดการขนส่งสร้างการเชื่อมต่อสำหรับคู่ คิว

โหมดการขนส่งต่อไปนี้ที่ได้รับการสนับสนุน

- Reliable connection (RC)
  - แต่ละ queue pair (QP) ที่สัมพันธ์กับ QP อื่น
  - ข้อความที่ถูกส่งโดยคิวการส่งของ QP หนึ่งที่ เชื่อถือได้ถูกนำส่งไปยังคิวรับของอีก QP
  - แพ็กเก็ตที่ถูกส่งตามลำดับ
  - RC คล้ายกับการเชื่อมต่อ TCP
- Unreliable datagram (UD)
  - การเชื่อมต่อที่ไม่มีจริงถูกจัดรูปแบบระหว่าง QP
  - โหมด UD คล้ายกับการเชื่อมต่อ UDP

การวางแผนสำหรับ Open Fabrics enterprise Distribution (OFED)

ไฟล์คอนฟิกูเรชันต้องมีอยู่ในไดเร็กทอรี /etc/libibverbs.d/ สำหรับทุกอะแด็ปเตอร์ Remote Direct Memory Access (RDMA) ที่ติดตั้ง บนระบบ

ไฟล์คอนฟิกูเรชันเปิดให้ไลบรารี **libibverbs** สามารถ ใช้ไดร์เวอร์สำหรับอุปกรณ์ RDMA ตัวอย่างเช่น เมื่อต้องการใช้อะแด็ปเตอร์ **Mellanox ConnectX-2 RoCE** ไฟล์ `mx2.driver` ต้องอยู่ในไดเร็กทอรี `/etc/libibverbs.d/` ไฟล์ `mx2.driver` ต้องมีโค้ดต่อไปนี้:

```
# cat /etc/libibverbs.d/mx2.driver
driver mx2
```

เมื่อต้องการใช้ไดเร็กทอรีอื่น ยกเว้นไดเร็กทอรี `/etc/libibverbs.d/` ใช้ตัวแปรสภาวะแวดล้อม `IBV_CONFIG_DIR` เมื่อต้องการสร้าง การสื่อสารระหว่างสองโหนด อะแด็ปเตอร์ต้องมี IPv4 หรือ IPv6 addresses กำหนดคอนฟิก

การสร้างการเชื่อมต่อโดยใช้ตัวจัดการการสื่อสาร (RDMA_CM)

ตัวจัดการการสื่อสาร Remote Direct Memory Access (RDMA) `RDMA_CM` มีการจัดการการสื่อสารที่รวมการตั้งค่า และ tear down การเชื่อมต่อสำหรับ RDMA application programming interface (API)

ตัวจัดการการสื่อสาร `RDMA_CM` ทำงานกับ verbs API ที่กำหนดโดยไลบรารี **libibverbs** ไลบรารี **libibverbs** จะมีอินเตอร์เฟสที่จำเป็นสำหรับส่งและรับข้อมูล

การดำเนินการไคลเอ็นต์

เรียนรู้เกี่ยวกับภาพรวมของการดำเนินการพื้นฐานสำหรับ การสื่อสารที่แอ็คทีฟ หรือไคลเอ็นต์

โฟลว์การเชื่อมต่อทั่วไปเป็นดังนี้:

rdma_create_event_channel
: สร้างแชนเนลเพื่อรับเหตุการณ์

rdma_create_id
: จัดสรรตัวบ่งชี้ `rdma_cm_id` ที่โดยหลักการแล้ว คล้ายกับซ็อกเก็ต

rdma_resolve_addr
: จัดหารอุปกรณ์ Remote Direct memory Access (RDMA) รีโมตเพื่อไปถึง รีโมตแอดเดรส

rdma_get_cm_event
: รอเหตุการณ์ `RDMA_CM_EVENT_ADDR_RESOLVED`

rdma_ack_cm_event
: ตอบรับเหตุการณ์ที่ได้รับ

rdma_create_qp
: จัดสรร queue pair (QP) สำหรับการสื่อสาร

rdma_resolve_route
: พิจารณาเส้นทางไปยังรีโมตแอดเดรส

rdma_get_cm_event
: รอเหตุการณ์ `RDMA_CM_EVENT_ROUTE_RESOLVED`

rdma_ack_cm_event
: ตอบรับเหตุการณ์ที่ได้รับ

rdma_connect
เชื่อมต่อกับรีโมตเซิร์ฟเวอร์

rdma_get_cm_event
รอเหตุการณ์ RDMA_CM_EVENT_ESTABLISHED

rdma_ack_cm_event
ตอบรับเหตุการณ์ที่ได้รับ

ibv_post_send()
ดำเนินกการถ่ายโอนข้อมูลผ่านการเชื่อมต่อ

rdma_disconnect
แบ่งส่วนการเชื่อมต่อ

rdma_get_cm_event
รอเหตุการณ์ RDMA_CM_EVENT_DISCONNECTED

rdma_ack_cm_event
ตอบกลับเหตุการณ์

rdma_destroy_qp
ทำลาย QP

rdma_destroy_id
รีลีสตัวบ่งชี้ rdma_cm_id

rdma_destroy_event_channel
รีลีสแชนเนลเหตุการณ์

หมายเหตุ: ในตัวอย่าง ไคลเอ็นต์เริ่มต้นการยกเลิกการเชื่อมต่อ อย่างไรก็ตาม กระบวนการไคลเอ็นต์ หรือเซิร์ฟเวอร์อย่างใดอย่างหนึ่งสามารถเริ่มต้นกระบวนการ ยกเลิกการเชื่อมต่อ

การดำเนินการของเซิร์ฟเวอร์

เรียนรู้เกี่ยวกับการดำเนินการพื้นฐานที่สามารถรันสำหรับ การสื่อสาร passive หรือเซิร์ฟเวอร์

โฟลว์การเชื่อมต่อทั่วไปเป็นดังนี้:

rdma_create_event_channel
สร้างแชนเนลเพื่อรับเหตุการณ์

rdma_create_id
จัดสรรตัวบ่งชี้ rdma_cm_id ที่โดยหลักการแล้ว คล้ายกับซ็อกเก็ต

rdma_bind_addr
ตั้งค่าหมายเลขพอร์ตโลคัลซึ่งเหตุการณ์ listens

rdma_listen
เริ่มทำการ listen การร้องขอการเชื่อมต่อ

rdma_get_cm_event
รอเหตุการณ์ RDMA_CM_EVENT_CONNECT_REQUEST ที่มีตัวบ่งชี้ rdma_cm_id

rdma_create_qp
จัดสรร queue pair (QP) สำหรับการสื่อสารกับตัวบ่งชี้ rdma_cm_id ใหม่

rdma_accept
ยอมรับการร้องขอการเชื่อมต่อ

rdma_ack_cm_event
ตอบกลับเหตุการณ์

rdma_get_cm_event
รองเหตุการณ์ RDMA_CM_EVENT_ESTABLISHED

rdma_ack_cm_event
ตอบกลับเหตุการณ์

ibv_post_send()
ดำเนินการถ่ายโอนข้อมูลผ่านการเชื่อมต่อ

rdma_get_cm_event
รอเหตุการณ์ RDMA_CM_EVENT_DISCONNECTED

rdma_ack_cm_event
ตอบกลับเหตุการณ์

rdma_disconnect
แบ่งส่วนการเชื่อมต่อ

rdma_destroy_qp
ทำลาย QP

rdma_destroy_id
รีลีสตัวบ่งชี้ **rdma_cm_id** ที่เชื่อมต่อ

rdma_destroy_id
รีลีสตัวบ่งชี้ rdma_cm_id ที่กำลัง listen

rdma_destroy_event_channel
รีลีสแชนเนลเหตุการณ์

ตัวอย่างตัวจัดการการสื่อสาร RDMA_CM

เรียนรู้เกี่ยวกับตัวอย่างที่แสดงชุมชน Open Fabrics Enterprise Distribution (OFED) ระหว่างการประชุม LinuxConf.Europe 2007

ข้อมูลที่เกี่ยวข้อง:

ตัวอย่างที่นำเสนอในชุมชน OFED

ตัวอย่างของแอ็คทีฟไคลเอ็นต์

ตัวอย่างของการดำเนินการการสื่อสารที่ไคลเอ็นต์ แอ็คทีฟ

```
/*
 * build:
 * cc -o client client.c -lrdmacm -libverbs
 *
 * usage:
 * client <servername> <val1> <val2>
 *
 * connects to server, sends val1 via RDMA write and val2 via send,
 * and receives val1+val2 back from the server.
 */
#include        <stdio.h>
#include <stdlib.h>
#include <stdint.h>
#include  <string.h>
#include <sys/types.h>
#include <sys/socket.h>
#include <netdb.h>
#include <arpa/inet.h>

#include <rdma/rdma cma.h>
enum   {
        RESOLVE TIMEOUT MS        = 5000,
};
struct pdata {
        uint64_t buf va;
        uint32_t buf rkey;
};

int main(int argc, char   *argv[ ])
{
   struct pdata         *server pdata;
   struct rdma_event channel  *cm_channel;
   struct rdma_cm_id      *cm_id;
   struct rdma_cm_event     *event;
   struct rdma_conn_param    conn_param = { };
   struct ibv_pd        *pd;
   struct ibv_comp_channel   *comp_chan;
   struct ibv_cq        *cq;
   struct ibv_cq        *evt_cq;
   struct ibv_mr        *mr;
   struct ibv_qp_init_attr   qp attr = { };
   struct ibv_sge        sge;
   struct ibv_send_wr      send_wr = { };
   struct ibv_send_wr      *bad send wr;
   struct ibv_recv_wr      recv wr = { };
   struct ibv_recv_wr      *bad recv wr;
   struct ibv_wc        wc;
   void             *cq context;
   struct addrinfo       *res, *t;
   struct addrinfo       hints = { .ai_family    = AF INET,
                           .ai_socktype  = SOCK STREAM
                                       };
```

```
int              n;
uint32_t          *buf;
int err;

      /* Set up RDMA CM structures */
  cm_channel = rdma_create_event_channel();
  if (!cm_channel)  return 1;
  err = rdma_create_id(cm_channel, &cm_id, NULL, RDMA_PS_TCP);
  if (err)
    return err;
  n = getaddrinfo(argv[1], "20079", &hints, &res);
  if (n < 0)
    return 1;

  /* Resolve server address and route */
  for (t = res; t; t = t->ai next) {
    err = rdma_resolve_addr(cm_id, NULL, t->ai_addr, RESOLVE_TIMEOUT_MS);
    if (!err)
      break;
  }
  if (err)
    return err;
  err = rdma_get_cm_event(cm_channel, &event);
  if (err)
    return err;
  if (event->event != RDMA_CM_EVENT_ADDR_RESOLVED)
    return 1;
  rdma_ack_cm_event(event);
  err = rdma_resolve_route(cm_id, RESOLVE_TIMEOUT_MS);
  if (err)
    return err;
  err = rdma_get_cm_event(cm_channel, &event);
  if (err)
    return err;
  if (event->event != RDMA_CM_EVENT_ROUTE_RESOLVED)
    return 1;
  rdma_ack_cm_event(event);

  /* Create verbs objects now that we know which device to use */
  pd = ibv_alloc_pd(cm_id->verbs);
  if (!pd)
    return 1;
  comp chan = ibv_create_comp_channel(cm_id->verbs);
  if (!comp_chan)
    return 1;
  cq = ibv_create_cq(cm_id->verbs, 2,NULL, comp_chan, 0);
  if (!cq)
    return 1;
   if (ibv_req_notify_cq(cq, 0))
    return 1;
   buf = calloc(2, sizeof (uint32_t));
  if (!buf)
    return 1;
  mr = ibv_reg_mr(pd, buf,2 * sizeof(uint32_t), IBV_ACCESS_LOCAL_ WRITE);
  if (!mr)
```

```
  return 1;
qp_attr.cap.max        send_wr = 2;
qp_attr.cap.max        send_sge = 1;
qp_attr.cap.max        recv_wr = 1;
qp_attr.cap.max        recv_sge = 1;
qp_attr.send_cq        = cq;
qp_attr.recv_cq        = cq;
qp_attr.qp_type        = IBV_QPT_RC;
err = rdma_create_qp(cm_id, pd, &qp_attr);
if (err)
  return err;
conn_param.initiator_depth = 1;
conn_param.retry_count     = 7;

/* Connect to server */
err = rdma_connect(cm_id, &conn_param);
if (err)
  return err;
err = rdma_get_cm_event(cm_channel,&event);
if (err)
  return err;
if (event->event != RDMA_CM_EVENT_ESTABLISHED)
  return 1;
memcpy(&server_pdata, event->param.conn.private_data, sizeof server_pdata);
rdma_ack_cm_event(event);

/* Prepost receive */
sge.addr                      = (uintptr_t) buf;
sge.length                    = sizeof (uint32_t);
sge.lkey                      = mr->lkey;
recv_wr.wr_id =     0;
recv_wr.sg_list =   &sge;
recv_wr.num_sge =   1;

if (ibv_post_recv(cm_id->qp, &recv_wr, &bad_recv_wr))
  return 1;

/* Write/send two integers to be added */
buf[0] = strtoul(argv[2], NULL, 0);
buf[1] = strtoul(argv[3], NULL, 0);
printf("%d + %d = ", buf[0], buf[1]);
buf[0] = htonl(buf[0]);
buf[1] = htonl(buf[1]);

sge.addr                      = (uintptr_t) buf;
sge.length                    = sizeof (uint32_t);
sge.lkey                      = mr->lkey;
send_wr.wr_id                 = 1;
sendwr.opcode                 = IBV_WR_RDMA_WRITE;
send _wr.sg_list              = &sge;
send_wr.num_sge               = 1;
send_wr.wr.rdma.rkey          = ntohl(server_pdata.buf_rkey);
send_wr.wr.rdma.remote_addr   = ntohll(server_pdata.buf_va);

  if (ibv post send(cm id->qp, &send_wr,  &bad_send_wr))
```

```
  return 1;
  sge.addr                    = (uintptr_t) buf + sizeof (uint32_t);
  sge.length                  = sizeof (uint32_t);
  sge.lkey                    = mr->lkey;
  send_wr.wr_id               = 2;
  send_wr.opcode              = IBV_WR_SEND;
  send_wr.send_flags          = IBV_SEND_SIGNALED;
  send_wr.sg_list             =&sge;
  send_wr.num_sge             = 1;

 if (ibv_post_send(cm_id->qp, &send_wr,&bad_send_wr))
  return 1;

  /* Wait for receive completion */
  while (1) {
     if (ibv_get_cq_event(comp_chan,&evt_cq, &cq_context))
    return 1;
  if (ibv_req_notify_cq(cq, 0))
    return 1;
  if (ibv_poll_cq(cq, 1, &wc) != 1)
    return 1;
  if (wc.status != IBV_WC_SUCCESS)
    return 1;
  if (wc.wr_id == 0) {
    printf("%d\n", ntohl(buf[0]));
    return 0;
  }
   }
   return 0;
 }
```

คำสั่ง OFED

เรียนรู้เกี่ยวกับคำสั่ง Open Fabrics Enterprise Distribution (OFED) รวมถึงข้อความสั่งของไวยากรณ์ คำอธิบายแฟล็ก และตัวอย่างการใช้งาน

คำสั่ง ibv_devices

แสดงรายการอุปกรณ์ Remote Direct Memory Access (RDMA) ที่มี สำหรับใช้งานจากพื้นที่ผู้ใช้

คำสั่ง ibv_devinfo

พิมพ์ข้อมูลเกี่ยวกับอุปกรณ์ RDMA network interface controller (RNIC) ที่มีให้ใช้จากพื้นที่ผู้ใช้

ไวยากรณ์

```
ibv_devinfo [-v] { [-d <dev>] [-i <port>] } | [-l]
```

แฟล็ก

รายการ	คำอธิบาย
-d *dev*	ใช้อุปกรณ์ *dev* RDMA โดยดีฟอลต์ อุปกรณ์แรกที่พบจะถูกใช้
-i *port*	ใช้พอร์ต *port* ของอุปกรณ์ RDMA โดย ดีฟอลต์ พอร์ตทั้งหมดถูกใช้
-l	พิมพ์ชื่ออุปกรณ์ RDMA เท่านั้น
-v	พิมพ์แอ็ตทริบิวต์ทั้งหมดของอุปกรณ์ RDMA

คำสั่ง ofedctrl

โหลด และเลิกโหลดส่วนขยายเคอร์เนล **ofed_core**

ไวยากรณ์

ofedctrl { [-k *KernextName*] -l|u|q }|[-c|-p *ParameterName=Value*]|-h

แฟล็ก

รายการ	คำอธิบาย
-c	รีโหลดไฟล์คอนฟิกูเรชันถ้าไฟล์ ถูกแก้ไข
-h	ระบุการใช้งาน
-k *KernextName*	ระบุพาธส่วนขยายเคอร์เนล โดยดีฟอลต์ พาธ /usr/lib/drivers/ofed_core ถูกใช้
-l	โหลดส่วนขยายเคอร์เนล
-p *ParameterName=Value*	ตั้งค่าของพารามิเตอร์โดยตรงบน บรรทัดรับคำสั่ง **หมายเหตุ:** ค่าที่ตั้งโดยใช้อ็อพชัน -p จะไม่ persistent. อ็อพชัน -p เปลี่ยนแปลง การกำหนดคอนฟิกปัจจุบันเท่านั้น จะไม่อัพเดตไฟล์คอนฟิกูเรชัน การเปลี่ยนแปลงที่ทำโดยใช้อ็อพชัน -p ไม่ มีผลใช้หลังจากระบบรีสตาร์ท
-q	ระบุว่าส่วนขยายเคอร์เนลถูกโหลด หรือไม่
-u	ยกเลิกการโหลดส่วนขยายเคอร์เนล

คำสั่ง rping

ทดสอบการเชื่อมต่อของตัวจัดการการสื่อสาร RDMA (RDMA_CM) โดยใช้การทดสอบ RDMA ping-pong

ไวยากรณ์

```
rping -s [-v] [-V] [-d] [-P] [-a address] [-p port] [-C message_count] [-S message_size]
rping -c [-v] [-V] [-d] -a address [-p port] [-C message_count] [-S message_size]
```

คำอธิบาย

คำสั่ง **rping** สร้างการเชื่อมต่อ Remote Direct Memory Access (RDMA) ที่น่าเชื่อถือ ระหว่างสองโหนดโดยใช้ไลบรารี **librdmacm** ทางเลือก คำสั่ง **rping** ยังดำเนินการถ่ายโอน RDMA ระหว่างโหนด จากนั้นยกเลิกการเชื่อมต่อ คำสั่ง **rping** ตั้งค่าการเชื่อมต่อ RDMA_CM และดำเนินการทดสอบ RDMA ping-pong สำหรับ ข้อมูลเกี่ยวกับคำสั่ง **rping** ดูที่ Open Source OpenFabrics Alliance OFED 1.4 ที่ http://www.openfabrics.org

แฟล็ก

รายการ	คำอธิบาย
-a *address*	ระบุแอดเดรสเครือข่ายเพื่อโยงการเชื่อมต่อ บนเซิร์ฟเวอร์ และระบุเซิร์ฟเวอร์แอดเดรสเพื่อเชื่อมต่อกับไคลเอ็นต์
-c	รันเป็นไคลเอ็นต์
-C *message_count*	ระบุจำนวนข้อความเพื่อถ่ายโอน ผ่านแต่ละการเชื่อมต่อ ค่าดีฟอลต์คือไม่จำกัด
-d	แสดงข้อมูลการดีบัก
-p	ระบุหมายเลขพอร์ตสำหรับเซิร์ฟเวอร์ ที่กำลัง listen
-P	รันเซิร์ฟเวอร์ในโหมดถาวร โหมดนี้อนุญาต ให้หลายไคลเอ็นต์ **rping** เชื่อมต่อกับอินสแตนซ์เซิร์ฟเวอร์เดียว และเซิร์ฟเวอร์รันจนกระทั่งเซิร์ฟเวอร์ถูกลบทิ้ง
-v	แสดงข้อมูล ping
-V	ตรวจสอบความถูกต้องข้อมูล ping
-s	รันเป็นเซิร์ฟเวอร์
-S *message_size*	ระบุขนาดของแต่ละข้อความที่ถ่ายโอน เป็นไบต์ ค่าดีฟอลต์คือ 100

ข้อมูลที่เกี่ยวข้อง:

Openfabrics

User-level Direct Access Programming Library (uDAPL)

User Direct Access Programming Library (uDAPL) คือเฟรมเวิร์ก การเข้าถึงโดยตรงที่รันบนการขนส่งที่สนับสนุนการเข้าถึงข้อมูลโดยตรง เช่น InfiniBand และ RDMA network interface controller (RNIC)

DAT Collaborative ระบุ uDAPL application programming interface (API) uDAPL codebase จาก Open Fabrics ถูกพอร์ตไปยังระบบปฏิบัติการ AIX และ ได้รับการสนับสนุนบน GX++ HCA และ 4X DDR Expansion card (CFFh) InfiniBand adapters

หลักการที่เกี่ยวข้อง:

"uDAPL APIs ที่สนับสนุนในระบบปฏิบัติการ AIX" ในหน้า 15
User Direct Access Programming Library (uDAPL) APIs ที่ระบุโดย DAT Collaborative ทั้งหมดไม่ได้รับการสนับสนุน โดยระบบปฏิบัติการ AIX
"แอ็ตทริบิวต์ที่ระบุเฉพาะผู้ขายสำหรับ uDAPL" ในหน้า 16
เรียนรู้เกี่ยวกับแอ็ตทริบิวต์ที่ผู้ขายเจาะจงที่ได้รับการ สนับสนุนโดยระบบปฏิบัติการ AIX โดยแอ็ตทริบิวต์ `delayed_ack_supported`, `vendor_extension`, `vendor_ext_version`, `debug_query` และ `debug_modify` จะได้รับการสนับสนุน

ข้อมูลที่เกี่ยวข้อง:

Datcollaborative

การติดตั้ง uDAPL

User-level Direct Access Programming Library (uDAPL) เวอร์ชัน 2.0 ได้รับการสนับสนุนโดยระบบปฏิบัติการ AIX

อิมเมจการติดตั้ง uDAPL ถูกส่งมาในแพ็กเสริมเป็น `udapl.rte` อิมเมจส่งไฟล์ส่วนหัว DAT ซึ่งอยู่ในไดเร็กทอรี `/usr/include/dat` อิมเมจการติดตั้งยังส่งไลบรารี `libdat.a` และ `libdapl.a`

แอ็พพลิเคชันมีไฟล์ส่วนหัว DAT และลิงก์ที่มีไลบรารี `libdat.a` DAT ในไดเร็กทอรี `/usr/include/dat` เลเยอร์ DAT กำหนดไลบรารีการขนส่งโดยเฉพาะที่กำหนด อย่างเหมาะสม

ผู้ให้บริการ AIX uDAPL รีจิสเตอร์ตนเองกับรีจีสทรี DAT โดยใช้รายการไฟล์ dat.conf ไฟล์ /etc/dat.conf ถูกส่งมากับรายกากรดีฟอลต์ และไฟล์มีรายละเอียดเกี่ยวกับรูปแบบของรายการ

ไลบรารี uDAPL สนับสนุนการติดตามระบบ AIX สำหรับการดีบัก เหตุการณ์ การติดตามระบบ uDAPL เชื่อมต่อ ID ที่มี 5C3 (สำหรับ เหตุการณ์ DAPL), 5C4 (สำหรับเหตุการณ์ข้อผิดพลาด DAPL), 5C7 (สำหรับเหตุการณ์ DAT) และ 5C8 (สำหรับเหตุการณ์ข้อผิดพลาด DAT) ระดับการติดตามเริ่มต้นถูกแก้ไข โดยใช้ตัวแปรสภาวะแวดล้อม DAT_TRACE_LEVEL และ DAPL_TRACE_LEVEL ตัวแปรสภาวะแวดล้อมเหล่านี้ยอมรับค่าในช่วง 0 – 10 จำนวนเหตุการณ์ และจำนวนข้อมูลที่ติดตามจะเพิ่มขึ้น ด้วยระดับการติดตามหลักดังนี้:

```
TRC_LVL_ERROR  = 1
TRC_LVL_NORMAL = 3
TRC_LVL_DETAIL = 7
```

คุณลักษณะความสามารถในการให้บริการ AIX มาตรฐานอื่นๆ เช่นล็อกข้อผิดพลาด AIX จะถูกใช้เพื่อระบุปัญหาเมื่อติดตามเหตุการณ์ คุณลักษณะ ความสามารถในการให้บริการของเลเยอร์การขนส่งที่กำหนด เช่นคำสั่ง **ibstat** และการติดตามคอมโพเนนต์ InfiniBand ยังเป็นประโยชน์สำหรับการวิเคราะห์ ปัญหา

DAT APIs ส่งคืนอโค้ดส่งคืนมาตรฐานที่สามารถถอดรหัสโดยใช้ไฟล์ /usr/include/dat/dat_error.h คำอธิบายโดยละเอียดเกี่ยวกับโค้ดส่งคืนมีใน ข้อมูลจำเพาะ uDAPL จาก DAT Collaborative

uDAPL APIs ที่สนับสนุนในระบบปฏิบัติการ AIX

User Direct Access Programming Library (uDAPL) APIs ที่ระบุโดย DAT Collaborative ทั้งหมดไม่ได้รับการสนับสนุน โดยระบบปฏิบัติการ AIX

APIs ต่อไปนี้ได้รับการสนับสนุนโดยการใช้ uDAPL อุตสาหกรรมทั่วไป และสนับสนุนโดยระบบปฏิบัติการ AIX

APIs ต่อไปนี้ไม่ได้รับการสนับสนุนโดยการใช้ uDAPL อุตสาหกรรมทั่วไป และไม่ได้รับการสนับสนุนโดยระบบปฏิบัติการ AIX

API	เวอร์ชัน
dat_cr_handoff	// In DAT 2.0
dat_ep_create_with_srq	// In DAT 2.0
dat_ep_recv_query	// In DAT 2.0
dat_ep_set_watermark	// In DAT 2.0
dat_srq_create	// In DAT 2.0
dat_srq_post_recv	// In DAT 2.0
dat_srq_resize	// In DAT 2.0
dat_srq_set_lw	// In DAT 2.0
dat_srq_free	// In DAT 2.0
dat_srq_query	// In DAT 2.0

APIs เพิ่มเติมต่อไปนี้ที่ระบบปฏิบัติการ AIX ไม่สนับสนุน:

- dat_lmr_sync_rdma_read
- dat_lmr_sync_rdma_write
- dat_registry_add_provider
- dat_registry_add_provider

สำหรับ APIs ที่ไม่สนับสนุนทั้งหมด ระบบปฏิบัติการ AIX เป็นไปตาม กลไกเฉพาะที่อธิบายในข้อมูลจำเพาะ DAT เพื่อบ่งชี้รายการ API ที่ไม่สนับสนุน ซึ่งรวมค่าแอ็ตทริบิวต์ `max_srq` ที่เป็นศูนย์ และโค้ดส่งคืน `DAT_MODEL_NOT_SUPPORTED` เฉพาะเจาะจง ทั้งนี้ตามการนำไปใช้ในอุตสาหกรรม และข้อมูลจำเพาะ DAT โค้ด `DAT_NOT_IMPLEMENTED` สามารถถูกส่งคืนสำหรับฟังก์ชัน ซึ่งไม่ได้รับการสนับสนุน

การสนับสนุนสำหรับ API ที่เกี่ยวกับ remote memory region (RMR) เช่น *dat_rmr_create*, *dat_rmr_bind*, *dat_rmr_free* และ *dat_rmr_query* ขึ้นอยู่กับความสามารถ host channel adapter (HCA) ที่กำหนด และความสำเร็จหรือล้มเหลวจะถูกพิจารณาโดยเฟรมเวิร์ก InfiniBand ที่กำหนด ขณะนี้อะแด็ปเตอร์ GX++ HCA และ 4X DDR Expansion card (CFFh) InfiniBand ไม่สนับสนุน การดำเนินการ RMR

หลักการที่เกี่ยวข้อง:

"User-level Direct Access Programming Library (uDAPL)" ในหน้า 14
User Direct Access Programming Library (uDAPL) คือเฟรมเวิร์ก การเข้าถึงโดยตรงที่รันบนการขนส่งที่สนับสนุนการเข้าถึงข้อมูลโดยตรง เช่น InfiniBand และ RDMA network interface controller (RNIC)

"แอ็ตทริบิวต์ที่ระบุเฉพาะผู้ขายสำหรับ uDAPL"
เรียนรู้เกี่ยวกับแอ็ตทริบิวต์ที่ผู้ขายเจาะจงที่ได้รับการ สนับสนุนโดยระบบปฏิบัติการ AIX โดยแอ็ตทริบิวต์ `delayed_ack_supported`, `vendor_extension`, `vendor_ext_version`, `debug_query` และ `debug_modify` จะได้รับการสนับสนุน

ข้อมูลที่เกี่ยวข้อง:


uDAPL: User Direct Access Programming Library

แอ็ตทริบิวต์ที่ระบุเฉพาะผู้ขายสำหรับ uDAPL

เรียนรู้เกี่ยวกับแอ็ตทริบิวต์ที่ผู้ขายเจาะจงที่ได้รับการ สนับสนุนโดยระบบปฏิบัติการ AIX โดยแอ็ตทริบิวต์ `delayed_ack_supported`, `vendor_extension`, `vendor_ext_version`, `debug_query` และ `debug_modify` จะได้รับการสนับสนุน

ผู้ให้บริการ AIX สำหรับการส่งผ่าน InfiniBand (IB) ประกอบด้วยแอ็ตทริบิวต์ interface adapter (IA) ที่ระบุเฉพาะผู้ขายซึ่งมีชื่อว่า `delayed_ack_supported` ค่าของแอ็ตทริบิวต์ `delayed_ack_supported` เป็น **`true`** หรือ **false** เมื่อค่าเป็น **`true`** จุดปลายที่ สัมพันธ์กับ IA® มี แอ็ตทริบิวต์ `delayed_ack` ที่ผู้ให้บริการเจาะจงที่แก้ไขได้ เมื่อแอ็ตทริบิวต์ `delayed_ack_supported` เป็น **false** จุดปลายของแอ็ตทริบิวต์ `delayed_ack` ที่ผู้ให้บริการเจาะจง จะไม่สามารถเปลี่ยนแปลงได้ ค่าดีฟอลต์ของจุดปลายของแอ็ตทริบิวต์ `delayed_ack` ที่ผู้ให้บริการเจาะจง คือ **false** แอ็ตทริบิวต์ `delayed_ack` ถูกตั้งค่าเป็น **`true`** โดยใช้อ็อพชัน `dat_ep_modify` ที่เปิดใช้งานคุณลักษณะการตอบรับที่หน่วงเวลาของ underlying InfiniBand (IB) host channel adapter (HCA) สำหรับคู่คิว InfiniBand ที่เจาะจงที่สัมพันธ์กับจุดปลาย คุณลักษณะฮาร์ดแวร์นี้ ไม่ได้ถูกนไปใช้โดย HCAs ทั้งหมด ดังนั้นจึงใช้ได้กับบาง IA เท่านั้น เมื่อคุณลักษณะนี้ถูกเปิดใช้งาน การตอบรับที่ส่งโดย HCA จะถูกหน่วงเวลาจนกว่าจะพบการดำเนินการถ่ายโอนข้อมูลในหน่วยความจำ ระบบของเซิร์ฟเวอร์ กระบวนการนี้ทำให้เวลาแฝงขนาดเล็กเพิ่มขึ้น

สำหรับการดีบักข้อผิดพลาด ไลบรารี uDAPL สนับสนุนการตามรอยระบบ AIX ระดับ การติดตามเริ่มต้นสามารถเปลี่ยนแปลงได้โดยใช้ตัวแปรสภาวะแวดล้อม `DAT_TRACE_LEVEL` and `DAPL_TRACE_LEVEL` เมื่อต้องการเปลี่ยนแปลงระดับการติดตามเหล่านี้แบบไดนามิกโดยใช้ API ให้ใช้การสนับสนุนระดับการติดตามไดนามิกบน AIX เมื่อต้องการตรวจสอบว่า ไลบรารีมีการสนันสนุนระดับการติดตามไดนามิกหรือไม่ แอ็พพลิเคชันสามารถเคียวรี แอ็ตทริบิวต์ IA `vendor_extension` ที่ผู้ขายเจาะจงได้ การแสดงตนของแอ็ตทริบิวต์ `vendor_extension` บ่งชี้ ระดับการติดตามไดนามิกที่ได้รับการสนับสนุน เมื่อแอ็ตทริ

บิวต์ vendor_extension ถูกแสดง แอ็พพลิเคชันสามารถเข้าถึงฟังก์ชันพอยเตอร์ dat_trclvl_query() และ dat_trclvl_modify() โดยการเคียวรีแอ็ตทริบิวต์ IA ที่ผู้ขายเจาะจงคือ debug_query และ debug_modify ค่าของแอ็ตทริบิวต์ เหล่านี้ชี้ไปยังฟังก์ชันที่เกี่ยวข้อง เมื่อต้องการทำให้อินเตอร์เฟส vendor_extension นี้พร้อมใช้งานในอนาคต ต้องใช้แอ็ตทริบิวต์ IA ที่ผู้ขายเจาะจงคือ vendor_extension ขณะนี้ แอ็ตทริบิวต์ vendor_extension ถูกตั้งค่าเป็น 1.0 และเป็นเวอร์ชันนี้เท่านั้นที่ได้รับการสนับสนุน ถ้าแอ็ตทริบิวต์ vendor_extension ไม่มีอยู่ แอ็พพลิเคชันจะไม่สามารถแก้ไขระดับการติดตามแบบไดนามิก

ตัวอย่างของวิธีเปลี่ยนแปลงแอ็ตทริบิวต์เหล่านี้มีอยู่ใน โค้ดตัวอย่าง uDAPL ที่ติดตั้งมากับการนำ AIX ไปใช้

หลักการที่เกี่ยวข้อง:

"uDAPL APIs ที่สนับสนุนในระบบปฏิบัติการ AIX" ในหน้า 15

User Direct Access Programming Library (uDAPL) APIs ที่ระบุโดย DAT Collaborative ทั้งหมดไม่ได้รับการสนับสนุน โดยระบบปฏิบัติการ AIX

"User-level Direct Access Programming Library (uDAPL)" ในหน้า 14

User Direct Access Programming Library (uDAPL) คือเฟรมเวิร์ก การเข้าถึงโดยตรงที่รันบนการขนส่งที่สนับสนุนการเข้าถึงข้อมูลโดยตรง เช่น InfiniBand และ RDMA network interface controller (RNIC)

คำประกาศ

ข้อมูลนี้จัดทำขึ้นสำหรับผลิตภัณฑ์และเซอร์วิสที่นำเสนอในสหรัฐฯ

IBM อาจไม่นำเสนอผลิตภัณฑ์ เซอร์วิส หรือคุณลักษณะที่อธิบายในเอกสารนี้ในประเทศอื่น โปรดปรึกษาตัวแทน IBM ในท้องถิ่นของคุณสำหรับข้อมูลเกี่ยวกับผลิตภัณฑ์ และเซอร์วิส ที่มีอยู่ในพื้นที่ของคุณในปัจจุบัน การอ้างอิงใดๆ ถึงผลิตภัณฑ์ โปรแกรม หรือเซอร์วิสของ IBM ไม่ได้มีวัตถุประสงค์ที่จะระบุหรือตีความว่า สามารถใช้ได้เฉพาะผลิตภัณฑ์ โปรแกรม หรือเซอร์วิสของ IBM เพียงอย่างเดียว เท่านั้น ผลิตภัณฑ์ โปรแกรม หรือเซอร์วิสใดๆ ที่สามารถทำงานได้เท่าเทียมกัน และไม่ละเมิดสิทธิทรัพย์สินทางปัญญาของ IBM อาจนำมาใช้ แทนได้ อย่างไรก็ตาม ถือเป็นความรับผิดชอบของผู้ใช้ที่จะประเมินและตรวจสอบการดำเนินการของ ผลิตภัณฑ์ โปรแกรม หรือเซอร์วิสใดๆ ที่ไม่ใช่ของ IBM

IBM อาจมีสิทธิบัตร หรืออยู่ระหว่างดำเนินการขอ สิทธิบัตรที่ครอบคลุมถึงหัวข้อซึ่งอธิบายในเอกสารนี้ การนำเสนอเอกสารนี้ไม่ได้เป็นการให้ไลเซนส์ใดๆ ในสิทธิบัตรเหล่านี้แก่คุณ คุณสามารถส่งการสอบถามเกี่ยวกับไลเซนส์ เป็นลายลักษณ์อักษรไปยัง:

IBM Director of Licensing
IBM Corporation
North Castle Drive, MD-NC119
Armonk, NY 10504-1785
United States of America

หากมีคำถามเกี่ยวกับข้อมูลชุดอักขระไบต์คู่ (DBCS) โปรดติดต่อแผนกทรัพย์สินทางปัญญาของ IBM ในประเทศของคุณ หรือส่งคำถาม เป็นลายลักษณ์อักษร ไปยัง:

Intellectual Property Licensing
Legal and Intellectual Property Law
IBM Japan Ltd.
19-21, Nihonbashi-Hakozakicho, Chuo-ku
Tokyo 103-8510, Japan

ย่อหน้าต่อไปนี้ไม่ได้ใช้กับสหราชอาณาจักร หรือประเทศอื่นใดที่ ข้อกำหนดดังกล่าวไม่สอดคล้องกับกฎหมายท้องถิ่น: INTERNATIONAL BUSINESS MACHINES CORPORATION นำเสนอสิ่งพิมพ์นี้ "ตามสภาพ" โดยไม่มีการรับประกันใดๆ โดยชัดแจ้งหรือ โดยนัย ซึ่งรวมถึงแต่ไม่จำกัดเฉพาะการรับประกันโดยนัยถึงการไม่ละเมิด การขายได้ หรือความเหมาะสมสำหรับวัตถุประสงค์เฉพาะ เนื่องจากบางรัฐไม่อนุญาตให้ปฏิเสธการรับประกันโดยชัดแจ้งหรือ โดยนัยในธุรกรรมบางอย่าง ดังนั้น ข้อความสั่งนี้จึงอาจไม่ใช้กับคุณ

ข้อมูลนี้อาจมีความไม่ถูกต้องด้านเทคนิคหรือข้อผิดพลาดจากการพิมพ์ มีการเปลี่ยนแปลง ข้อมูลในเอกสารนี้เป็นระยะ และการเปลี่ยนแปลงเหล่านี้จะรวมอยู่ในเอดิชันใหม่ของ สิ่งพิมพ์ IBM อาจปรับปรุง และ/หรือเปลี่ยนแปลงในผลิตภัณฑ์ และ/หรือโปรแกรมที่อธิบายในสิ่งพิมพ์นี้ได้ตลอดเวลา โดยไม่ต้องแจ้งให้ทราบ

การอ้างอิงใดๆ ในข้อมูลนี้ถึงเว็บไซต์ไม่ใช่ของ IBM มีการจัดเตรียมเพื่อความสะดวกเท่านั้น และ ไม่ได้เป็นการรับรองเว็บไซต์เหล่านั้นในลักษณะใดๆ เอกสารประกอบที่เว็บไซต์เหล่านั้นไม่ได้เป็น ส่วนหนึ่งของเอกสารประกอบสำหรับผลิตภัณฑ์ IBM นี้ และการใช้เว็บไซต์เหล่านั้นถือเป็นความเสี่ยงของคุณเอง

IBM อาจใช้ หรือแจกจ่าย ข้อมูลใดๆ ที่คุณให้ในวิธีที่ IBM เชื่อว่าเหมาะสมโดยไม่ก่อให้เกิดข้อผูกมัดใดๆ กับ คุณ

ผู้รับไลเซนส์ของโปรแกรมนี้ที่ต้องการข้อมูลเกี่ยวกับโปรแกรมสำหรับวัตถุประสงค์ในการเปิดใช้งาน: (i) การแลกเปลี่ยนข้อมูลระหว่างโปรแกรมที่สร้างขึ้นอย่างอิสระกับโปรแกรมอื่น (รวมถึง โปรแกรมนี้) และ (ii) การใช้ข้อมูลซึ่งแลกเปลี่ยนร่วมกัน ควร ติดต่อ:

IBM Corporation
Dept. LRAS/Bldg. 903
11501 Burnet Road
Austin, TX 78758-3400
USA

ข้อมูลดังกล่าวอาจพร้อมใช้งาน ภายใต้ข้อตกลงและเงื่อนไขที่เหมาะสม รวมถึง การชำระค่าธรรมเนียมในบางกรณี

โปรแกรมที่มีไลเซนส์ซึ่งอธิบายในเอกสารนี้ และเอกสารประกอบที่มีไลเซนส์ทั้งหมดสำหรับโปรแกรม นั้น มีการจัดเตรียมโดย IBM ภายใต้ข้อตกลงของข้อตกลงกับลูกค้าของ IBM, ข้อตกลงไลเซนส์โปรแกรมระหว่างประเทศของ IBM หรือข้อตกลงที่เท่าเทียมกันใดๆ ระหว่างเรา

ข้อมูลประสิทธิภาพใดๆ ที่มีในเอกสารนี้ถูกกำหนดในสภาวะแวดล้อมที่ควบคุม ด้วยเหตุนี้ ผลลัพธ์ที่ได้ในสภาวะแวดล้อมการปฏิบัติการอื่นจึงอาจแตกต่างไปอย่างมาก การวัดบางอย่างอาจ ดำเนินการบนระบบที่อยู่ระหว่างการพัฒนา และไม่มีการรับประกันว่าการวัดเหล่านี้จะ เหมือนกันบนระบบที่พร้อมใช้งานโดยทั่วไป ยิ่งไปกว่านั้น การวัดบางอย่างอาจมีการประเมินโดยวิธีการ ประมาณค่านอกช่วง ผลลัพธ์จริงอาจแตกต่างไป ผู้ใช้เอกสารนี้จึงควรตรวจสอบ ข้อมูลที่สามารถใช้ได้สำหรับสภาวะแวดล้อมของตน

ข้อมูลเกี่ยวกับผลิตภัณฑ์ที่ไม่ใช่ของ IBM ได้รับมาจากซัพพลายเออร์ของผลิตภัณฑ์เหล่านั้น ประกาศที่เผยแพร่ หรือแหล่งข้อมูลที่เปิดเผยต่อสาธารณะ IBM ไม่ได้ทดสอบผลิตภัณฑ์ดังกล่าว และไม่สามารถยืนยันความถูกต้องของ ประสิทธิภาพ ความเข้ากันได้ หรือการเรียกร้องอื่นใดที่เกี่ยวข้องกับผลิตภัณฑ์ที่ไม่ใช่ของ IBM คำถามเกี่ยวกับ ความสามารถของผลิตภัณฑ์ที่ไม่ใช่ของ IBM ควรส่งไปยังซัพพลายเออร์ของผลิตภัณฑ์เหล่านั้น

ข้อความทั้งหมดเกี่ยวกับทิศทางหรือเจตนาในอนาคตของ IBM อาจมีการเปลี่ยนแปลง หรือเพิกถอนได้โดยไม่ต้องแจ้งให้ทราบ และแสดงถึงเป้าหมายและวัตถุประสงค์เท่านั้น

ราคาของ IBM ทั้งหมดที่แสดงเป็นราคาขายปลีกที่แนะนำของ IBM ซึ่งเป็นราคาปัจจุบัน และอาจเปลี่ยนแปลงได้โดยไม่ต้องแจ้งให้ทราบ ราคาของผู้แทนจำหน่ายอาจแตกต่างไป

ข้อมูลนี้ใช้สำหรับวัตถุประสงค์ของการวางแผนเท่านั้น ข้อมูลในเอกสารนี้อาจมีการเปลี่ยนแปลง ก่อนผลิตภัณฑ์ที่อธิบายจะวางจำหน่าย

ข้อมูลนี้มีตัวอย่างของข้อมูลและรายงานที่ใช้ในการดำเนินการทางธุรกิจรายวัน เพื่อ สาธิตข้อมูลให้สมบูรณ์ที่สุดเท่าที่จะเป็นไปได้ ตัวอย่างจึงมีชื่อของแต่ละบุคคล บริษัท ยี่ห้อ และผลิตภัณฑ์ ชื่อทั้งหมดเหล่านี้เป็นชื่อสมมติ และการคล้ายคลึงในชื่อและที่อยู่ซึ่งหน่วยธุรกิจจริงใช้เป็นความบังเอิญโดยสิ้นเชิง

ไลเซนส์ลิขสิทธิ์:

ข้อมูลนี้มีตัวอย่างแอ็พพลิเคชันโปรแกรมในภาษาต้นฉบับ ซึ่งแสดงถึง เทคนิคด้านโปรแกรมในหลากหลายแพล็ตฟอร์ม คุณอาจคัดลอก ปรับเปลี่ยน และแจกจ่าย โปรแกรมตัวอย่างเหล่านี้ในรูปแบบใดๆ โดยไม่ต้องชำระเงินให้แก่ IBM สำหรับวัตถุประสงค์ในการพัฒนา การใช้ การตลาด หรือการแจกจ่ายโปรแกรมแอ็พพลิเคชัน ที่สอดคล้องกับอินเตอร์เฟสการเขียนโปรแกรมแอ็พพลิเคชันสำหรับแพล็ตฟอร์มปฏิบัติการ ซึ่งเขียน โปรแกรมตัวอย่าง ตัวอย่างเหล่านี้ยังไม่ได้ผ่านการทดสอบในทุกสภาพ ดังนั้น IBM จึงไม่สามารถรับประกัน หรือบอกเป็นนัยถึง ความน่าเชื่อถือ ความสามารถบริการได้ หรือฟังก์ชันของโปรแกรมเหล่านี้ โปรแกรมตัวอย่างมีการนำเสนอ "ตาม สภาพ" โดยไม่มีการรับประกันประเภทใดๆ IBM ไม่รับผิดชอบ ต่อความเสียหายใดๆ ที่เกิดขึ้นเนื่องจากการใช้โปรแกรมตัวอย่างของคุณ

แต่ละสำเนา หรือส่วนใดๆ ของโปรแกรมตัวอย่างเหล่านี้ หรืองานที่สืบเนื่องใดๆ ต้องมี คำประกาศลิขสิทธิ์ดังนี้:

ส่วนของโค้ดนี้ได้มาจากโปรแกรมตัวอย่างของ IBM Corp.

สิ่งที่ต้องพิจารณาเกี่ยวกับนโยบายความเป็นส่วนตัว

ผลิตภัณฑ์ซอฟต์แวร์ของ IBM® รวมถึงโซลูชันบริการระบบซอฟแวร์ ("ข้อเสนอซอฟต์แวร์") อาจใช้คุกกี้หรือเทคโนโลยีอื่นเพื่อรวบรวมข้อมูลการใช้งานผลิตภัณฑ์ เพื่อช่วยในการปรับปรุงประสบการณ์การใช้งานของผู้ใช้ขั้นปลาย เพื่อปรับแต่งการโต้ตอบกับ ผู้ใช้ขั้นปลาย หรือเพื่อวัตถุประสงค์อื่นๆ ในหลายๆ กรณี จะไม่มีการรวบรวม ข้อมูลอัตลักษณ์ส่วนบุคคลโดย ข้อเสนอซอฟต์แวร์ ซึ่งข้อเสนอซอฟต์แวร์บางอย่าง สามารถช่วยให้คุณรวบรวมข้อมูลอัตลักษณ์ ส่วนบุคคลได้ ถ้าข้อเสนอซอฟต์แวร์นี้ใช้คุกกี้ เพื่อรวบรวมข้อมูลอัตลักษณ์, ระบุข้อมูล เกี่ยวกับการใช้คุกกี้ของข้อเสนอนี้ถูกกำหนดไว้ด้านล่าง

ข้อเสนอซอฟต์แวร์นี้ไม่ใช้คุกกี้ หรือเทคโนโลยีอื่นเพื่อรวบรวมข้อมูลอัตลักษณ์ส่วนบุคคล

ถ้าคอนฟิกูเรชันถูกปรับใช้สำหรับ ข้อเสนอกนี้จัดเตรียมให้คุณในฐานะลูกค้าสามารถรวบรวม ข้อมูลอัตลักษณ์ส่วนบุคคลจากผู้ใช้ขั้นปลายผ่านทางคุกกี้ และเทคโนโลยีอื่น คุณควรปรึกษากับที่ปรึกษาด้านกฎหมายเกี่ยวกับ ที่ใช้บังคับในการรวบรวมข้อมูล รวมถึงข้อกำหนดต่างๆ เพื่อการแจ้งเตือนและการยินยอม

สำหรับข้อมูลเพิ่มเติมเกี่ยวกับการใช้ เทคโนโลยีต่างๆ รวมถึงคุกกี้ สำหรับวัตถุประสงค์เหล่านี้ โปรดดู นโยบายความเป็นส่วนตัวของ IBM ที่ http://www.ibm.com/privacy และ คำชี้แจงสิทธิส่วนบุคคลออนไลน์ของ IBM ที่ส่วน http://www.ibm.com/privacy/details "Cookies, Web Beacons and Other Technologies" และ "IBM Software Products and Software-as-a-Service Privacy Statement" ที่ http://www.ibm.com/software/info/product-privacy

---

เครื่องหมายการค้า

IBM, ตราสัญลักษณ์ IBM , และ ibm.com เป็นเครื่องหมายการค้าหรือเครื่องหมายการค้าที่จดทะเบียนของ International Business Machines Corp. ซึ่งจดทะเบียนในหลายเขตอำนาจศาลทั่วโลก ชื่อผลิตภัณฑ์และการบริการอื่นอาจเป็นเครื่องหมายการค้าของ IBM หรือบริษัทอื่น รายการปัจจุบันของเครื่องหมายการค้า IBM มีอยู่บนเว็บไซต์ที่ ข้อมูลลิขสิทธิ์และเครื่องหมายการค้า ที่ www.ibm.com/legal/copytrade.shtml

INFINIBAND, InfiniBand Trade Association, และ ลักษณะแบบ INFINIBAND คือเครื่องหมายการค้าและ/หรือลักษณะเซอร์วิสของ INFINIBAND Trade Association

Linux เป็นเครื่องหมายการค้าจดทะเบียนของ Linus Torvalds ในสหรัฐอเมริกา ประเทศอื่นๆ หรือทั้งสองกรณี

ดัชนี

อักขระพิเศษ

A

O

R

U

V

ก

ค

ต

ล

ห

อ

พิมพ์ในสหรัฐอเมริกา