空运磁带的 PB 级实验数据传输

续本达

清华大学 工程物理系

2022-10-08 Tunight

Outline

磁盘阵列不是备份

磁带也不是备份

任务

XMASS 是世界上唯一的单相液态氙暗物质实验。在暗物质研究之外，液态氙中的 ^{124}Xe, ^{126}Xe 的双中微子 ($2\nu\epsilon\epsilon$) 和无中微子 ($0\nu\epsilon\epsilon$) 双电子俘获，以及 ^{134}Xe, ^{136}Xe 的双中微子 ($2\nu\beta\beta$) 和无中微子 ($0\nu\beta\beta$) 双贝塔衰等物理过程，是与中微子质量本源联系紧密的核物理现象。我对它们有浓厚兴趣。然而，实验附属的计算集群即将关闭，要推进下一步研究，必须把近 700TB 的原始实验数据传输到清华。

TL;DR

把 700TB 数据从日本传到清华

现有方案

- 台式机上的 LTO-7 驱动器，人工换带
  - 没人帮我，我也无法到现场

现有方案

- 台式机上的 LTO-7 驱动器，人工换带
  - 没人帮我，我也无法到现场
- 必须买全自动方案
  - HPE StoreEver MSL 3040 或者 Neos StorageLibrary T24 LTO8 或者 Qualstar Q24 LTO 8
  - 日本 Amazon 下单磁带库，无法顺利完成对公支付

现有方案

- 台式机上的 LTO-7 驱动器，人工换带
  - 没人帮我，我也无法到现场
- 必须买全自动方案
  - HPE StoreEver MSL 3040 或者 Neos StorageLibrary T24 LTO8 或者 Qualstar Q24 LTO 8
  - 日本 Amazon 下单磁带库，无法顺利完成对公支付
- 自己掏钱：6 万人民币
  - 美国厂商拒绝发到亚洲，买家退款
  - 直接联系美国厂商，洽谈到中间，被拒绝

```
Unfortunately we are unable to process this order, we have strict g
for shipping Qualstar Tape libraries internationally.
```

现有方案

- 台式机上的 LTO-7 驱动器，人工换带
  - 没人帮我，我也无法到现场
- 必须买全自动方案
  - HPE StoreEver MSL 3040 或者 Neos StorageLibrary T24 LTO8 或者 Qualstar Q24 LTO 8
  - 日本 Amazon 下单磁带库，无法顺利完成对公支付
- 自己掏钱：6 万人民币
  - 美国厂商拒绝发到亚洲，买家退款
  - 直接联系美国厂商，洽谈到中间，被拒绝

    ```
    Unfortunately we are unable to process this order, we have strict g
    for shipping Qualstar Tape libraries internationally.
    ```
- 联系国内厂商
  - 从美国进口，到达中国。在中国报关，发到日本。预计需要 4 个月时间。
  - 价格 10 万以上，约 1 万运费。

峰回路转 谢谢 reddit

- 磁带库的兼容性

  I've got a TL2000 and TL4000 library. Was running LTO5 FC drives in it. I went and bought a LTO8 HH drive listed as being for a TS4300.

  ...

  I put it in my TL2000 library, and initially the front panel showed the previous serial number and model number of the LTO5 drive. After around 5-10 minutes, it updated, and correctly showed the LTO8 drive and serial number.

- 磁带库的生产商

  The small libraries (up to 4U in size) sold by HP, IBM, Fujitsu, Dell, Quantum, Oracle/Sun/StorageTek, SpectraLogic, Overland and others are mostly made by a German company called BDT. Having a look at The BDT Storage site you'll be able to find a few pics of libraries OEMed for their partner companies. Considering their broad array of re-badged customer products, BDT seems to be very good at what they do.

- 推论：从 LTO-1 开始，磁带库的机械结构都是同一家公司生产且一样的。

BDT 机架磁带库的总上游生产商

https://www.bdt.de/storage

Idea!

- 日本、北京各购买一台 Dell TL2000 磁带库
- 买一台 LTO-8 磁带机
- 在北京调通，把磁带机发送到日本安装。
- 解决所有问题：
  1. Dell TL2000 旧机便宜
  2. 只需要发磁带机，不需要发磁带库，运费降低一个数量级

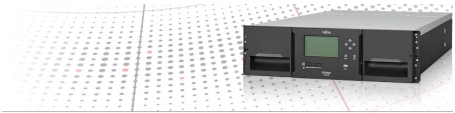
幸亏康哥买买买

- 康哥挖掘出了中国最强磁带机工程师
- 不需要在北京买 TL2000 了!
- 不需要调试磁带库和磁带机的兼容性了!
- 不需要买全新 LTO-8 磁带机了!
- Bonus 信息:
  - 只有 HP, Quantum 和 IBM 能生产 LTO 磁带机，其它都是贴牌
  - 只有 Fujitsu 和 SONY 能生产 LTO 磁带，其它都是贴牌

磁带库最终方案：两个多月的漫长摸索

穷人的磁带库：Dell TL2000(日本购买) + LTO-8 drive (发往日本)

奢侈的磁带库：ETERNUS LT140 with LTO-7 drive (买不起)



標準価格	314万2,000円（税別・最小構成時）より 【主要な構成内訳】LTO Ultrium7 ハーフハイトテープドライブ*1、収納巻数20巻

两个多月的漫长摸索

收获

感受到了从前科研中才有的“顿悟”快乐

原计划 15 盘 LTO-8

- xz 压缩率 4.5，700TB / 4.5 = 156TB
- zpaq 压缩率近一步提高
  - http://mattmahoney.net/dc/zpaq.html
  - 为 Windows 设计
- lrzip 调用 zpaq，压缩率更高，逼近 7
  - 9 盘 LTO-8 即可装下！

压缩测试

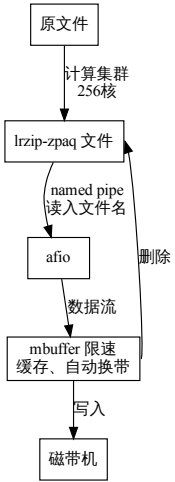
- 需要找一个压缩时间和比例的平衡

	time/s	threads	size/MB
xz -6	48	8	121
xz -6	374	1	120
xz -1	36	1	145
xz -3	101	1	144
zstd -1	2	1	220
zstd -10	23	1	174
zstd -19	341	1	140
lrzip -z -3	117	1	96
lrzip -z -4	386	1	91
lrzip -z -5	381	1	91
lrzip -l -7	24	8	282
lrzip -7	486	1	124
lrzip -z -7	138	8	92
lrzip -z -7	444	1	91
lrzip -z -9	335	8	89
lrzip -z -9	1036	1	88
lrzip -b -7	27	8	123
lrzip -g -7	26	8	184

资源调度计划

- 157MB/s 向磁带机写入。
  - 使用 `lrzip -z -3`，单线程写速度为 $96/117 = 0.82\,\mathrm{MB/s}$
  - 需要至少 $157 / 0.82 = 191$ 个线程同时压缩
- 742MB/s 从原始数据读 ($\times$ 5.5)，保守估计压缩率.
  - 11 天时间
- 1TB 周转硬盘空间

步骤



总体调度 GNU Make，shell 脚本

集群调度 PBS interactive 模式运行 lrzip

`afio` 把压缩后的文件收集起来，开 named pipe 读入文件列表

- 每隔两小时排序压缩好的文件，文件名列表写入 named pipe

`mbuffer` raw stream 写入磁带机

- 限速为 112MB/s （最终）
- speed matching，防止磁带机往复运动
- 自动换带，磁带写满 hook 调用

  `mtx -f /dev/sgX next`

硬带的 “RAID 0”

- 都是顺序写入，设备互不影响。
- 仍有各种问题出现，实际耗时 20 天。

Date and time	Status of Item	Location
Aug 26, 2022 2:30 PM	Posted	5061205
Aug 28, 2022 7:39 PM	Arrived at export office	Nagoya, Aichi
Aug 28, 2022 7:40 PM	Departed from export office	Nagoya, Aichi
Sep 1, 2022 11:01 AM	Arrived at destination import office	Beijing
Sep 1, 2022 11:01 AM	Presented to import customs	Beijing
Sep 1, 2022 2:37 PM	Held for customs inspection	BJSDJCGJHHJ
Sep 10, 2022 8:48 PM	Presented to import customs	Beijing
Sep 11, 2022 4:08 PM	Presented to import customs	Beijing
Sep 12, 2022 10:47 AM	Released from import customs	Beijing
Sep 12, 2022 10:52 AM	Departed from destination import office	Beijing
Sep 12, 2022 6:46 PM	Arrived at post office	100084
Sep 12, 2022 6:46 PM	Out for delivery	100084
Sep 12, 2022 6:50 PM	Delivery attempted	100084
Sep 12, 2022 7:34 PM	Held at delivery depot	100084
Sep 13, 2022 8:30 AM	Out for delivery	100084
Sep 13, 2022 10:21 AM	Delivered	

到清华了

- 有一些机械损坏

拆包

- 数据没有受到影响

空运的 PB 级数据传输

```
Storage Element 19:Full :VolumeTag=NB5267L8
Storage Element 20:Full :VolumeTag=NB5268L8
Storage Element 21:Full :VolumeTag=NB5269L8
Storage Element 22:Full :VolumeTag=NB5270L8
Storage Element 23:Full :VolumeTag=NB5271L8
Storage Element 24:Full :VolumeTag=NB5272L8
Storage Element 25:Full :VolumeTag=NB5273L8
Storage Element 26:Full :VolumeTag=NB5274L8
Storage Element 27:Full :VolumeTag=NB5275L8
Storage Element 28:Full :VolumeTag=NB5276L8
```

- 700TB / 40 days = 700 / 40 * 8 * 1024 / 86400 Gbps = 1.66Gbps