实验物理垃圾佬的高性能 GPU 集群

武益阳

2025 年 12 月 27 日

背景故事

前情提要

曾经，续老师为了备份、运输、存储 PB 级实验数据，做了系列 tunight:

然而，光存储是不行的，还需要大量的算力分析。正好续老师带领团队研发出了超强但超吃矩阵乘法算力的 FSMP 算法。

Why ROCm – 来自续老师的一封信 i

- nVidia 卡已经成为理财产品，价格大起大落。基于 nVidia 生态的“研究”成为有钱人的游戏，例如 11 月初的二手市场上 A100 被量化金融集团炒到了 20 万元一块。
- 如果研究者加入有钱人的游戏，为了研究的顺利就必须一直去搞钱。搞钱耗费的时间与精力，不如投入到尝试使用更开放的 GPU 计算方案。
- CUDA 是封闭的软件生态，其内部的工作原理我们无从得知。ROCm 挑战了这一范式，给出了在源代码层面兼容 CUDA 的自由软件 GPU 通用计算系统。
- 我们要能够理解 GPU 计算的步骤，是科学计算可复现性的基本要求。
- ROCm 让我们对 GPU 计算有更深刻的理解，也让我们降低对 vendor lock-in 技术与硬件的依赖，方便在必要时刻转至国产技术。

Why ROCm – 来自续老师的一封信 ii

- 而使用国产技术有助于降低申请国家级基金资助的难度。进一步让我们少花精力去搞钱，专注于研究本身。
- 自由软件的副产品是升级的过程更平滑，Linux 内核中有 ROCm 原生的支持。不需要学习另一套与 FHS 不符的额外约定。从远来看折腾更少。

搞 ROCm!

经过一个 OSPP-2021 + GSoC-2022，多个开发者的合作努力下，Gentoo ROCm 生态欣欣向荣！

图 1: OSPP2021 tunight

图 2: 你可以在市场上买到几千元的 MI100（图中价格仅供参考），计算性能逼近/超越 A100 (如通用 FP64)

GPU 集群需求

- **算力估计**：1 MI100 支撑 80 波形/s
- JUNO 探测器的子探测器 OSIRIS 已经开始运行，600 波形/s
- 2025 开始采数的 JUNO 探测器可达 40,000 波形/s
- 随着实验规模的扩大，对 GPU 数目的需求随之增多
- 如何利用有限的经费与场地部署尽可能多的 GPU？

高性能、高密度、高性价比 GPU 集群建设

垃圾佬的圣地：广州 i

垃圾佬的圣地：广州 ii

技嘉 G292-Z20 8卡GPU服务器AI人工智能 深度学习 4090卡 PCIE4.0

已售 31

¥3600

配　　送：广东佛山　至　北京 海淀

快递: 100.00
预计23小时内发货｜承诺48小时内发货

保　　障：7天无理由退货 极速退款

颜色分类：准系统 支持7001-7002 不含 CPU 内存…

准系统 支持7002-7003 不含 CPU 内存…

图 3: 你可以在佛山买到便宜的二手服务器

理想 vs 现实：G292-Z20 原型机 lilypad

- 理想：拼起来就是一套完美的机器
- 现实：拼起来一堆 bug
  - 商家忘刷 BIOS 不认 Zen3 CPU
  - 主板质量问题，有一通道内存无法识别
  - 内存体质不好，插上无法开机
  - 内存体质不好，经常 ECC 报错
  - 默认风扇配置根本压不住 GPU
  - 没有原装导轨，替代方案总不如意
    - 最终在帮物理系友军迁移机房的过程中发现 DELL C6420 静态导轨与 Z20 的 ** 几乎一样 **（只需替换螺丝即可）
    - 咸鱼入手 6 条导轨！
- 在各种 debug 过程中逐渐理解了大厂研发服务器背后的付出
- 使用高质量二手配件：合理的价格，站在大厂的肩膀上，加上自己的一点点适配

中途尝试的 G292-Z45：原型机 kaiserin i

- 缘起：为了导轨这瓶醋，买了 G292-Z45 这盘饺子
  - 由于台海局势，技嘉不再采用大陆工厂的导轨方案，转而采购台湾导轨
  - 拆机的 G292-Z20 莫名其妙都没有伴随导轨。要原厂导轨只能找原厂买
  - 技嘉：单买导轨运费太不划算了，请再买一个服务器吧

中途尝试的 G292-Z45：原型机 kaiserin ii

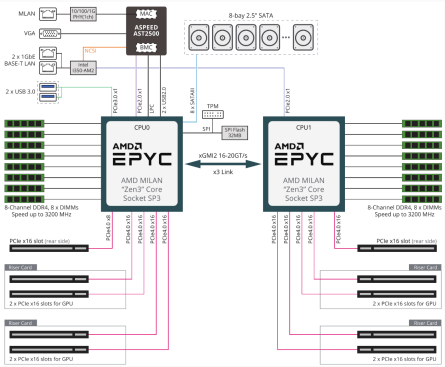


图 7: Z45 系统组成图：双 CPU，8 卡平台

中途尝试的 G292-Z45：原型机 kaiserin iii

图 8： Z45 实物图

中途尝试的 G292-Z45: 原型机 kaiserin iv

- 理论上 x2 CPU <-> GPU 带宽
- 实际 GPU 使用体验和 Z20 差不多
- 原厂价格 26000

最终方案: 6 台 G292-Z20 集群 ceres[0-5]

- 委托由手握大量 Z20 系统的商家
- 使用新 (生产出来就闲置着的) EPYC 7C13
- 同样爱好 ROCm 的量化基金老板掏出了库存的 MI100 (自己在用更具性价比的 MI50)
- 导轨怎么办？
  - 最终在帮物理系友军迁移机房的过程中发现 DELL C6420 静态导轨与 Z20 的 ** 几乎一样 **（只需替换螺丝即可）
  - 咸鱼入手 6 条导轨！

那么，捡 AI 垃圾代价是什么呢？电力与散热难题

服务器机房群微信群上的聊天记录如下，请查收。

————2025-09-20 ————

隔壁 LHCb 大佬运维 16:20

半个多小时前有人动了这个开关吗

我自己 16:23

半个小时之前我们有四台服务器莫名其妙关机了，不知道是否有关

我自己 16:24

从日志判断疑似是有一侧电源不在了

隔壁 LHCb 大佬运维 16:24

我们服务器也断电了，现在开回来了，你看看有没有起来

我自己 16:26

我们的服务器是从两排插座取电的，看起来是单排断电导致了高负载的机器供电不足而自动关机，但 BMC 还活着

- 机箱后部 GPU 日常吃高温尾气：

```
amdgpu 0000:48:00.0: amdgpu: WARN: GPU thermal throttling temperature reached, expect performance decrease. HBM.

amdgpu 0000:c5:00.0: amdgpu: WARN: GPU thermal throttling temperature reached, expect performance decrease. GPU.
```

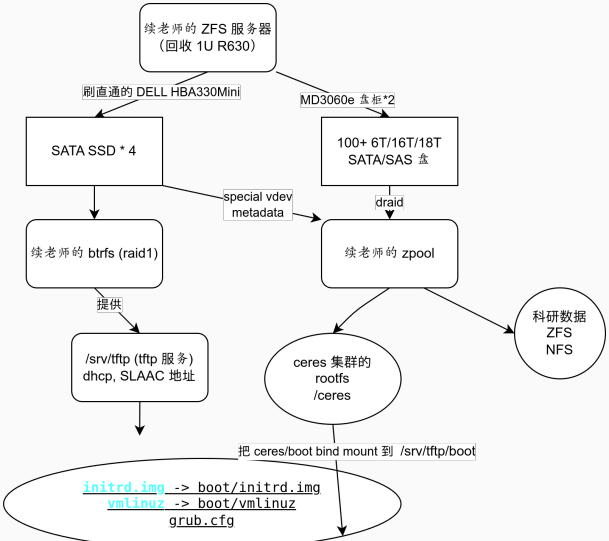
- 16500 RPM 暴力风扇的噪音轻易穿透机房门，响彻整个地下车库

慎入廉价二手 DDR4 3200 内存

- 内存体质不好，插上无法开机
- 内存体质不好，经常 ECC 报错
  - (2024) 价格 770 的三星 64G (M393A8G40AB2-CWE) 故障率 8/35，来回售后了三轮
  - (2024) 价格 950 的三星 64G (M393A8G40AB2-CWE) 故障率 1/46
  - (2024) 价格 300 的三星 32G (M393A4K40DB3-CWE) 故障率 1/8
- 隐约看出二手市场的层次化：价格差异反应在故障率上

灵车运维心得

openrc+NFS+tftp 无盘管理



- NFS 确保集群文件系统一致性
- 单独 /var/log NFS 避免日志冲突
- 单个服务器运行 apt 升级即可
- bind mount boot: 让内核升级也无需额外操作即可传递给 PXE 启动
  - 一定记得 bind mount 要在 zfs-mount 后面

未来

还有哪些显卡可以选择

- 继续探寻 MI100 渠道
- 量化基金老板的建议：MI50 16GB 性价比超高
- 对于希望用 NV 的小伙伴：V100 价格也不高
- 4090：尺寸、散热与 Z20 系统不兼容，需要改装 and 限制 TDP
- 瞄准 FP32 算力，魔改 AMD 民用卡？
  - 需要魔改电源接口、散热器：PCB 虽然勉强符合 PCIe 标准，但供电散热尺寸起飞
  - 散热器魔改可能需要开模

自古妖板出华擎

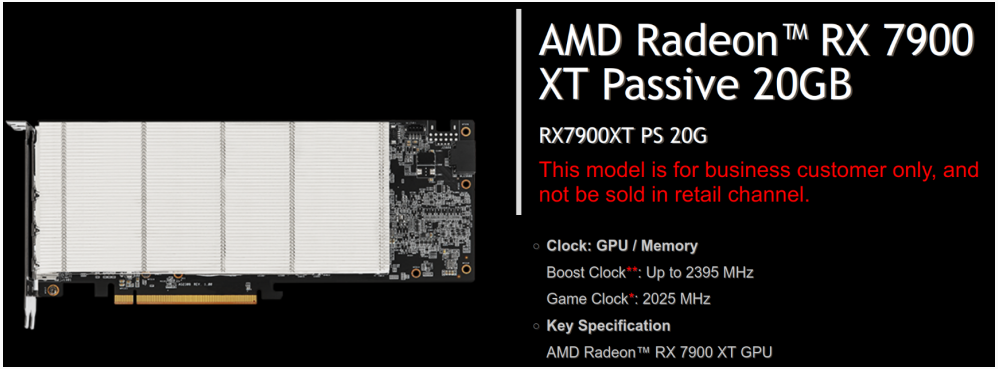


图 9: 7900XTX Passive

可惜深圳的一级代理表示尚未有交期，持续跟进中（估计已经凉了）

彩蛋

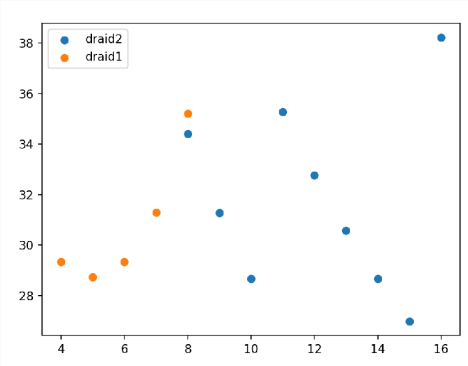
续老师的新硬盘：发现以前对 draid 理解不够透彻

https://github.com/openzfs/zfs/issues/13727#issuecomment-1278352959

draid ZFS 实际可用空间公式

```
# D: 一组内 data 盘数, P: 一组内parity
# 盘数, T: 总盘数, 不包含 hot spare
32 / (math.ceil(32/D) * D) * D/(D+P) * T
```

A	B	C	D	E	F	G	H	I	J	K
1	5	9	13	17	21	24	27	30	P1	P2
2	6	10	14	18	22	25	28	31	P1	P2
3	7	11	15	19	23	26	29	32	P1	P2
4	8	12	16	20	xx	xx	xx	xx	P1	P2



45 盘情况下不同 draid 的可用盘数

ambiance: 高性能塔式工作站

ambiance: 高性能塔式工作站

- 预算 5 万，交给上海的商家
- Gigabyte MZ72-HB2
- 2 * EPYC 7773X X3D 大三缓 (target OpenFOAM 流体力学)
- 16*32 GB 3200 内存（坏了一根）
- 2 * AMD 7900XTX
  - 曾部署 LLM 但没有 Motivation 去打理
  - 接入 slurm 但 GPU 用户体验不够好，慢慢没人用了
- 散热问题：塔式工作站放工作站内对附近人员进行热量攻击（曾遭到举报）
- 存在玄学网卡故障
- 独显接显示器会夺舍 AST2600 集显，导致 IPMI 远程 KVM 无显示

ambiance: 小作坊全闪

- 5 * SlimSAS 接口 + 2 * PCIe 4.0x8 -> 4 * SlimSAS (但没有热插拔) + 1 * M2 = 10 盘 1.6TB 全闪!
  - 一半 2T 消费级固态：三星 PM9A1，致态 Tipro7000，联想 OEM 长江存储 PC411 + PCIE 4.0 NVME 转 U2 硬盘盒 (70 RMB)
  - 捡企业级垃圾：华为 ES3600P V6，希捷 5350H
  - 10 盘 NVME U.2 硬盘笼 ()
  - 记得加装风扇散热！华为盘用了三个月坏一个（zfs 挺过来了）
  - 不要开 O_{DIRECT}!
- 显卡档 PCIe -> 使用显卡延长线!

Acknowledgements

GPU 集群

- 感谢国家自然科学基金与续本达课题组提供的经费支持
- 续老师打开的 ROCm 之路发掘 G292-Z20 二手供应链、采购协助、技术框架
- 课题组同学（含 Berrysoft）协助安装、汇报 bug

ambiance 塔式工作站

- 感谢国家自然科学基金与女朋友的经费支持
- 北师大相关同学的采购协助
- Adamanteye 协助安装全闪盘阵