RISC-V 软件移植及优化锦标赛 —— 第三次赛题讲解 & 演示

S2311：Baby LLaMA 2 on Duo 速度优化

（儿童讲故事场景）

https://rvspoc.org/s2311/

RVSPOC 组委会 – Ryan

2024 年 1 月 2 日

题目解析 S2311

项目描述：

让 Baby LLaMA 2 运行在 Milk-V Duo 这样的小板子上是很有挑战的事情。本次竞赛旨在提升 Baby LLaMA 2 在 Milk-V Duo 平台上的性能，目标是实现更高的每秒 Token 处理速度。参赛者需要运用轻量级技术和编译器优化策略，结合麦克风语音输入或命令行输入提示词等多种方式，开发一个能够讲故事的机器人 Demo。该 Demo 应通过扬声器进行输出，并可借鉴小米米兔讲故事机器人的原型设计。

产出及评分要求：

1. 评审标准将聚焦于正确性和性能两个方面，赛题给定相同输入，分别通过基准测试对参赛作品的正确性和性能打分。
2. 正确性评分使用参赛作品的输出和基准输出的差分测试结果衡量，从而反映出参赛作品优化技术对模型推理精度的影响。性能评分使用每秒钟计算的 Token 数量衡量，这直接反映出参赛作品的性能优化效果。
3. 文本转语音（TTS）部分单独计算时间。
4. 最终，组委会将根据参赛作品的正确性和性能的综合表现进行评分，两者将按照赛题评审委员会设定的加权比例计算出最终得分，得分最高的参赛者将获得胜利。

验证平台：Duo **64M**

知识产权及开源协议说明：

所有参赛结果要求开源，并提交至主办方指定仓库。参赛者（作者）持有作品的所有权。主办方鼓励参赛者将结果回馈贡献至 upstream。

编译/演示环境

- 编译环境
  - Gentoo Linux AMD64
  - 使用官方默认 Docker 镜像 milkvtech/milkv-duo:latest
    - 地址： https://hub.docker.com/r/milkvtech/milkv-duo （注意其他镜像地址同步问题）
    - 上层是 Ubuntu 22.04
- 演示环境
  - Darwin 23.2.0 ARM64

开始之前——Baby Llama 2 简介

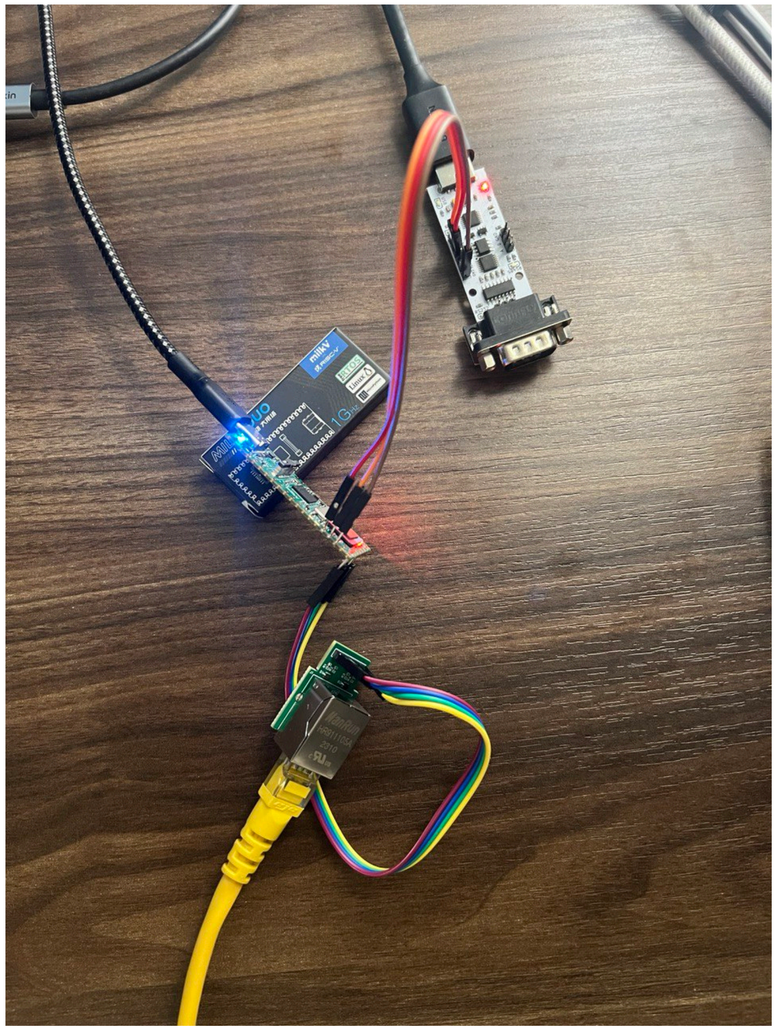
- 仓库：https://github.com/karpathy/llama2.c
- 一个实际代码不到 700 行且仅依赖标准库的 C 文件
- 兼容 Llama 2 模型（参数需小于 7B，由于其使用 32 位浮点数精度进行推断，速度很慢）
- 主要用于文本生成

开始之前——默认情景下的 Baby Llama 2 演示

- 使用官方预编译好的 Milk-V Duo 系统镜像：

  milkv-duo-v1.0.7-2023-1223.img.zip

- 使用该镜像默认的工具链编译 Baby Llama 2 的二进制。
- 硬件连接整体一览见右图
  - USB to TTL 用于更可靠的显示 Duo 的输出
  - RJ45 模块用于联网和传输数据



开始之前——默认情景下的 Baby Llama 2 演示

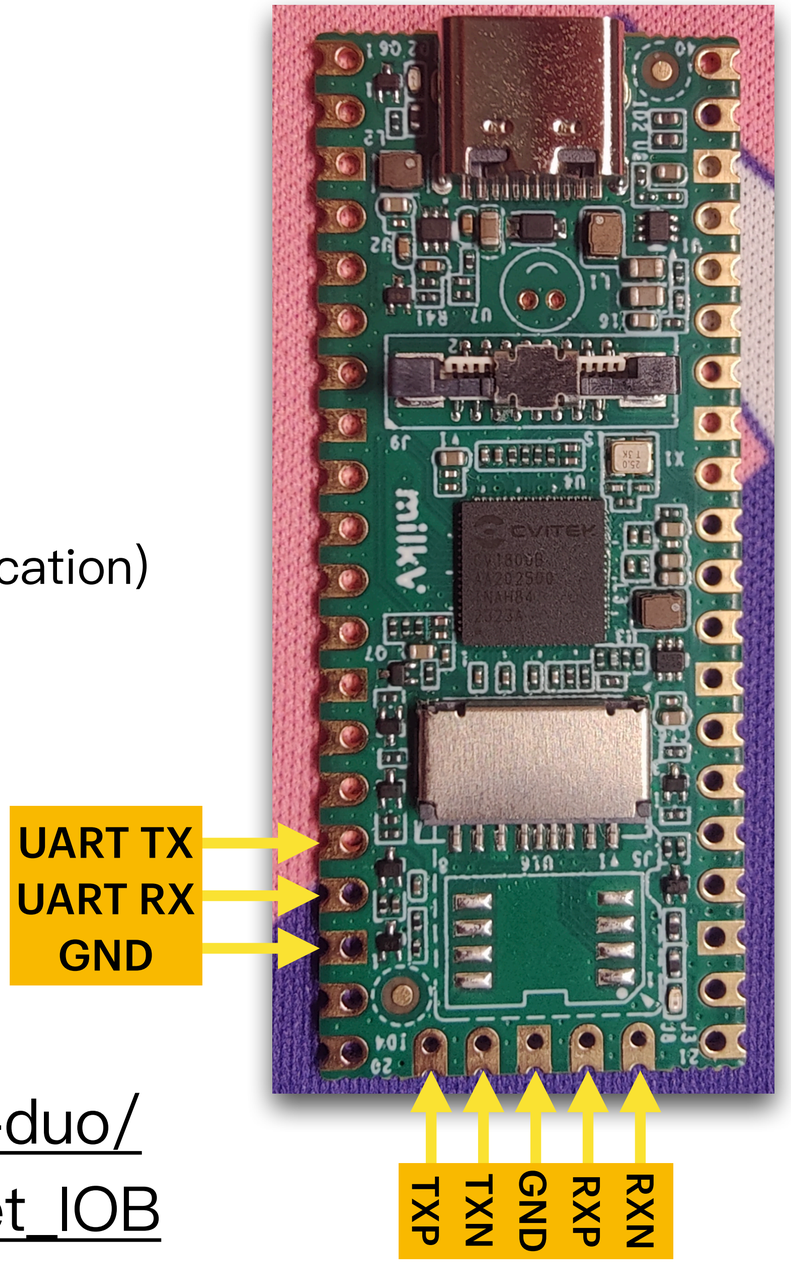
```
[root@milkv-duo]~# time ./run-gcc-musl-O3-rv64gcv0p7_zfh_xthead stories260K.bin
atepper con j F meriniaay h C ` $ S { iscode #es reigdees Com com sul Herameiten s hers =er      J wh lare com { anad j S aler p laqu Com
{ is m Winiaay s a to v th { isist thes re pro as coming withigdees Com com { is a wheson pella s ploow thct B { pro
 O ofra j (artistes w ` en rck jichraon p laqu Com {est Pel k thesle ` oldonle la enldersel      { Sor with ` kent      sent      {
 this Tdeterer comend j (ciing ret Der); pro o en to kst k hloub k jter {os k qu com dot rlest  peheined with plo ex k lare com on ramelee
rsas com unich S hilin laame enameic th jummentra j (//otoset D inoton hloub kin enellach comleed with s the
achieved tok/s: 187.224670
real    0m 1.37s
user    0m 1.19s
sys     0m 0.07s
[root@milkv-duo]~#
```

```
[root@milkv-duo]~# time ./run-gcc-musl-O3-rv64gcv0p7_zfh_xthead ./stories15M.bin
Once upon a time, there was a little girl named Lily. She had a sister named Rose. Lily and Rose liked to play in the park together. One d
ay, they saw a big tree with a swing hanging from it.
Lily wanted to go on the swing, but she was scared. Rose said, "Lily, let's ask Mom to take you on the swing." Lily agreed, and they went
to find Mom.
Mom took Lily and Rose on the swing. They laughed and had fun. After playing on the swing, they went home and told their mom about their a
dventure.
achieved tok/s: 0.293635
real    7m 36.53s
user    0m 18.11s
sys     0m 39.32s
```

如何部署此默认环境?

——接口介绍篇

- Milk-V Duo 的通讯方式：
  1. RNDIS (Remote Network Driver Interface Specification)
  2. 串口（TTL UART）
  3. 以太网物理接口
- 引脚说明：https://milkv.io/docs/duo/overview#gpio
- IO Board 原理图：https://github.com/milkv-duo/accessories/tree/master/Duo_USB&Ethernet_IOB



如何部署此默认环境?

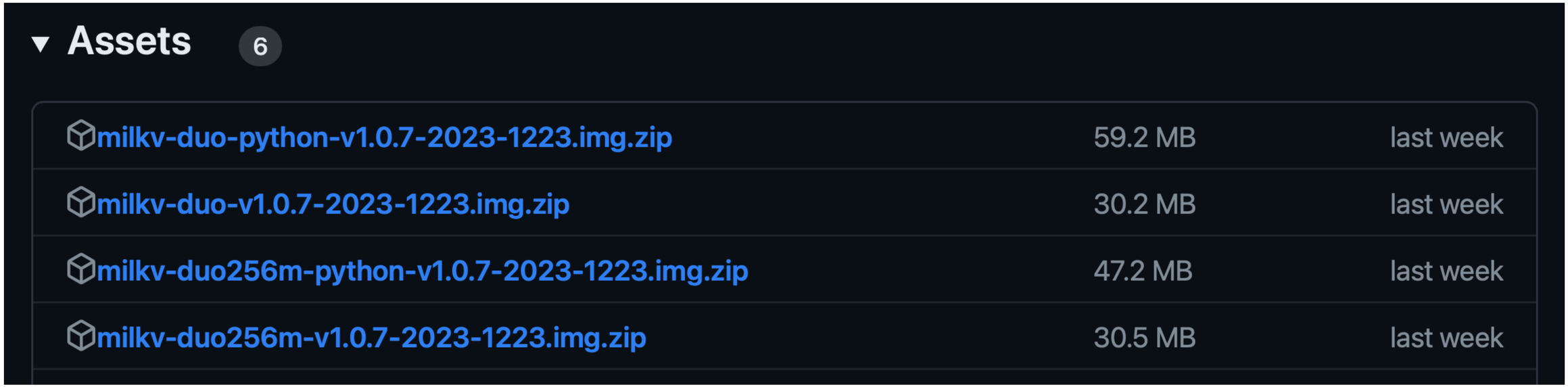
——接口选择篇

- 设定 Milk-V Duo 的通讯方式为：
  - 串口 & 以太网物理接口
- **串口**，这里支持使用 UART over TTL 方式，通常的 USB to TTL 芯片 CH340，CP2102 等之类的都可以支持。
- **以太网接口**，默认的 Milk-V Duo 是不带以太网接口的，但是它的芯片 CV1800b 支持 100Mpbs 的以太网传输，且预留了接口可用于扩展。比如官方给出的 IO Board 就扩展了。且原理图是开源的，这里我使用了现成的以太网接口扩展板。
- Q：为何要使用两个接口?
- Q：为何没有使用最方便的 RNDIS?

如何部署此默认环境?

——系统镜像准备篇

- 此默认环境使用的是 Milk-V 官方提供的默认镜像，直接下载即可。
  - 后文会有自定义系统镜像的说明。
- https://github.com/milkv-duo/duo-buildroot-sdk/releases



如何部署此默认环境?

—— TF 卡烧录篇

- 使用 `dd` 命令进行烧录

```
➜ # lsblk
➜ # ## dd 操作不可逆，注意选择到正确的设备，同时注意该设备内原有数据备份
➜ # dd if=/path/to/milkv-duo-v1.0.7-2023-1223.img of=/dev/sdX status=progress
➜ # sync
```

- `conv=sparse` ?
- 虽然镜像内实际的文件只有百兆左右，但由于在制作过程中对分区指定了大小，因此实际烧录过程中会写许多的「零」到 TF 卡上。后文会讲如何在自定义 rootfs 的时候避免这个问题，以及如何在烧录完 TF 卡后对其分区进行快速扩容。

如何部署此默认环境?

——启动

1. 将 TF 卡接入 Milk-V Duo
2. 使用杜邦线将 Milk-V Duo 上与 USB-to-TTL 的对应排针相连
3. 将 USB-to-TTL 设备接入演示环境
4. 将串口输出内容接入演示环境
5. 将 Milk-V Duo 设备接入演示环境，通电后，它将自动启动
6. 可以观察到演示环境下的串口输入内容显示 Milk-V Duo 的启动信息，最终停留在其 shell 环境中
7. 接入网线后可以尝试 ssh 到 Milk-V Duo 环境下

如何部署此默认环境?

—— Baby Llama 2 二进制准备篇

```
$ cd /path/to/duo-buildroot-sdk/dir #(optional)
$ git clone https://github.com/karpathy/llama2.c
$ cd llama2.c
$ ../host-tools/gcc/riscv64-linux-musl-x86_64/bin/riscv64-unknown-linux-musl-gcc -march=rv64gcv0p7_zfh_xtheadc -mabi=lp64d -mtune=c906 -O3 -lm -o run-gcc-musl-O3-rv64gcv0p7_zfh_xthead run.c
$ ../host-tools/gcc/riscv64-linux-musl-x86_64/bin/riscv64-unknown-linux-musl-gcc -march=rv64gcv0p7_zfh_xtheadc -mabi=lp64d -mtune=c906 -Ofast -lm -o run-gcc-musl-Ofast-rv64gcv0p7_zfh_xthead run.c
$ scp run-gcc-* root@<duo-ip>:~/
$ scp stories*.bin root@<duo-ip>:~/ #文件需要额外下载
$ scp tokenizer.bin root@<duo-ip>:~/
```

如何部署此默认环境?

——完结

至此，于最开始演示所需的默认环境就已经成立了。

如何对系统镜像进行自定义，做可行性调整?

——环境准备篇（the `milkv-duo:latest` Docker image）

- 官方目前声明的唯一兼容平台是在 Ubuntu 22.04 AMD64 上，于其他平台上做 rootfs 的话可能需要额外的精力来做适配，为了快速演示，这里使用官方发布的 Docker 镜像；本次使用支持 rootless 的 podman 作为 Docker 管理工具。
  - ✦ 文档：https://milkv.io/zh/docs/duo/getting-started/buildroot-sdk
  - ✦ 注意：pull 镜像时，某些 Linux 发行版可能默认指定了其他的（非 docker.io）镜像源，从而出现了类似 `manifest unknown` 这样的提示，这种情况一般是由于镜像未同步到而导致的，可以编辑 `/etc/containers/registries.conf` 进行修改。
- 步骤
  - ➡ `$ git clone --depth 1 https://github.com/milkv-duo/duo-buildroot-sdk.git /path/to/local/duo-buildroot-sdk/dir`
  - ➡ `$ podman pull milkvtech/milkv-duo:latest`
  - ➡ `$ podman run -it --rm -v /path/to/local/duo-buildroot-sdk/dir:/home/worker milkvtech/milkv-duo`

如何对系统镜像进行自定义，做可行性调整?

——系统配置篇之可用内存调整

- 为何明明有 64M 的物理内存，默认镜像显示的可用内存却仅有 28M 呢？这个问题在 buildroot-sdk 的 FAQ 里有提到（https://github.com/milkv-duo/duo-buildroot-sdk#faqs），是由于给 ION Buffer （用于 摄像头，Boot Logo 显示 等）预留了一部分。

  ✦ 可以通过修改 build/boards/cv180x/cv1800b_milkv_duo_sd/memmap.py 文件内 ION_SIZE 变量来修改为其预留的内存。

```
ION_SIZE = 26.80078125 * SIZE_1M
```

- 在 Baby Llama 2 针对 15M 参数量的模型进行推断的时候，占用内存很少，目前测试实际连几兆的内存都消耗不掉，故这里仅做参考。

如何对系统镜像进行自定义，做可行性调整?

——系统配置篇之分区调整

- 默认的分区总大小为 1152M，其中真实的数据仅占百兆左右，剩余空间均使用「零」进行了补足，如何根据需要来调整默认的分区呢?
  - 比如文件越丢越多，已经放不下了，或者 TF 需要有预留额外的分区给别的用处
- 这里将 boot 分区从 128M –> 8M
- 将默认的 root 分区从 768M –> 80M （单 milkv-duo 默认配置所需 root 分区大小为 60M+）
- 去掉未使用的预分配 256M Swap 分区
- 对应配置文件为：device/milkv-duo/genimage.cfg
  - 其使用的是 genimage 这个工具
  - 路径中的 milkv-duo 对应配置名称
- 对 ext4 文件系统进行扩容 (fdisk, resize2fs, e2fsck)（为了方便，提供一个额外的脚本）

```
image boot.vfat {
        vfat {
                label = "boot"
                files = {
                        "fip.bin",
                        "rawimages/boot.sd",
                }
        }
        size = 8M
}

image rootfs.ext4 {
        ext4 {
                label = "rootfs"
        }
        size = 80M
}

image milkv-duo.img {
        hdimage {
        }

        partition boot {
                partition-type = 0xC
                bootable = "true"
                image = "boot.vfat"
        }

        partition rootfs {
                partition-type = 0x83
                image = "rootfs.ext4"
        }
}
```

可能的 Baby Llama 2 性能优化点是哪些?

——编译篇

- Baby Llama 2 这个项目自带的 Makefile 文件就有多个编译目标
  - ✦ run –O3 优化
  - ✦ runfast –Ofast 优化
  - ✦ runomp 加入了 OpenMP 多线程优化（由于 Milk-V Duo 的 Linux 系统仅单核单线程，经测试使用此项会负优化）

(260K model)

- -O0: ~45 tok/s
- -O1: ~160
- -O2: ~186
- -O3: ~188
- -Ofast: ~200

```
[root@milkv-duo]~# ./run-gcc-musl-O3-rv64gcv0p7_zfh_xthead stories260K.bin
atepper con j F mer theit     ulersed {ionan k firerabalpectionlo exulersow {ionulersow m W a de sse forant trectaring {ionulersow m Wab s
al withst sun laia y {os k mab sleers not Com {
ort); withad jonulersowirer); proing withpe {os); coming withonulersow j F merab hameom ` {ionulersow mined withleers ` toub la `    ameer
she and kersor H {ionulersow mer  repe com lare jumonulersow m Winiaay sist thes hlo Dctesct {ion  re with proout laia en th siniaom th {
ue toram ofra j (ueptstla reg rendigonulersow unichionoram of sonulersow Ere inilin exel ther); Comiaersre {est a wh    peoweson dit it s
Pel k thesulersow
achieved tok/s: 188.330871
[root@milkv-duo]~# ./run-gcc-musl-Ofast-rv64gcv0p7_zfh_xthead ./stories260K.bin
atepper con j F mer " h C ` $ S { iscode #es re pro as coming with andde    peow Comel      { anad j S qu comterer comenddees w ` endeleie
la Bin la Gerar with `lo exloit comleielalo withag m     de k fes the withel);essant        nd com jumment m pameit      onde it `ub {
 O of qu comter aldeigde qu com a toame Com {os k merutelagmentoritenonleielaloere s Pou th     la com Com { pro - t); S p Tct =er    peow
 enpeow jmentlepereub la I Coman k ex t     er comdeter hpe Ddeerab d laction { isor with ` k     de k f s pameedic thes pameil with ex ks
e J nalersreri com {
 ha com you Ere rend j S alerab j proout lalo ex Comleers ` toub la `leielalo with { is mck that
achieved tok/s: 200.471698
```

可能的 Baby Llama 2 性能优化点是哪些?

——编译篇之二

```
16
17 # https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html
18 # https://simonbyrne.github.io/notes/fastmath/
19 # -Ofast enables all -O3 optimizations.
20 # Disregards strict standards compliance.
21 # It also enables optimizations that are not valid for all standard-compliant programs.
22 # It turns on -ffast-math, -fallow-store-data-races and the Fortran-specific
23 # -fstack-arrays, unless -fmax-stack-var-size is specified, and -fno-protect-parens.
24 # It turns off -fsemantic-interposition.
25 # In our specific application this is *probably* okay to use
26 .PHONY: runfast
27 runfast: run.c
```

- https://gcc.gnu.org/onlinedocs/gcc/Optimize-Options.html#Optimize-Options
- https://llvm.org/docs/Passes.html

```
➜ $ echo 'int;' | clang -xc -Ofast - -o /dev/null -\#\#\#
```

可能的 Baby Llama 2 性能优化点是哪些?

——编译篇之三 · 比较 1

AMD64 下

```
~/Git/llama2.c › make run
gcc -O3 -o run run.c -lm
gcc -O3 -o runq runq.c -lm
  1" .669439 +.088
  ::: SSH :::
~/Git/llama2.c › ./run stories15M.bin
One day, a boy named Tim went to the par
"Hi Sue!" said Tim. "Can I play with you
king the toy car go zoom.
After playing, they needed a magnet. The
toy car. Now, the car and the magnet wer
Tim and Sue laughed and played with the
achieved tok/s: 74.722459
  2" .394929 +.047
```

```
~/Git/llama2.c › make run CC=clang
clang -O3 -o run run.c -lm
clang -O3 -o runq runq.c -lm
  1" .272438 +.048
  ::: SSH :::
~/Git/llama2.c › ./run stories15M.bin
Once upon a time, there was a little gir
h her mom. They saw a big slide and Lily
"Mommy, can I go on the slide?" asked Li
"Sure, but first you have to lay down an
Lily laid down and closed her eyes. Her
Look, Mommy! A butterfly!" she exclaimed
Her mom smiled and said, "Yes, it's a pr
r's company.
achieved tok/s: 78.313253
  2" .540751 +.043
```

```
~/Git/llama2.c › make runfast
gcc -Ofast -o run run.c -lm
gcc -Ofast -o runq runq.c -lm
  1" .785686 +.097
  ::: SSH :::
~/Git/llama2.c › ./run stories15M.bin
Once upon a time, there was a little girl
 They walked through the foggy fields and
held onto her mommy's hand tightly.
"Mommy, what's that man saying?" asked Lil
"He's a security guard. He always makes su
Suddenly, the man's hat flew off his head
m. The man was very grateful and thanked t
As they continued their walk, Lily saw a b
 Her mommy and daddy were very worried and
achieved tok/s: 170.226969
  1" .537289 +.048
```

```
~/Git/llama2.c › make runfast CC=clang
clang -Ofast -o run run.c -lm
clang -Ofast -o runq runq.c -lm
  1" .284065 +.097
  ::: SSH :::
~/Git/llama2.c › ./run stories15M.bin
Once upon a time, there was a little girl
doll when she heard a loud noise outside.
Lily went outside to help her brother. Sh
 mommy and we'll get you a band-aid."
Lily got a band-aid and helped her brothe
. You can give it to my little brother as
Lily smiled and said, "Thank you, mommy.
achieved tok/s: 173.657718
  1" .231712 +.044
```

可能的 Baby Llama 2 性能优化点是哪些?

——编译篇之三 · 比较 2

ARM64 下

```
~/Git/llama2.c ➤ gcc --version
Apple clang version 15.0.0 (clang-1500.1.
Target: arm64-apple-darwin23.2.0
Thread model: posix
InstalledDir: /Library/Developer/CommandL
  .026358 +.043

~/Git/llama2.c ➤ make run
gcc -O3 -o run run.c -lm
gcc -O3 -o runq runq.c -lm
  .383691 +.036

~/Git/llama2.c ➤ ./run stories15M.bin
Once upon a time, there was a little girl
hat was very shiny and gold. It was a dia
Lily was so happy and wanted to show her
 asked her where she got it. Lily told th
At the end of the day, Lily went home and
the rock with her other special things in
achieved tok/s: 150.909091
  1" .823253 +.037
```

```
~/Git/llama2.c ➤ gcc --version
Apple clang version 15.0.0 (clang-1500.1.0.
Target: arm64-apple-darwin23.2.0
Thread model: posix
InstalledDir: /Library/Developer/CommandLin
  .024154 +.047

~/Git/llama2.c ➤ make runfast
gcc -Ofast -o run run.c -lm
gcc -Ofast -o runq runq.c -lm
  .412712 +.039

~/Git/llama2.c ➤ ./run stories15M.bin
One day, a little girl named Lily went to t
ird on the water. It had colorful feathers
Lily asked her mom, "Can I say hi to the bi
t to swim in the sea. She hoped she would s
Lily saw a small, cute fish in the water. S
ly picked it up and showed it to her mom. H
special, even if she didn't see the fancy b
achieved tok/s: 860.294118
  .834331 +.039
```

```
::: GP :::
~/Git/llama2.c ➤ gcc --version
gcc (Gentoo 13.2.0 p2) 13.2.0
Copyright (C) 2023 Free Software Foundati
This is free software; see the source for
warranty; not even for MERCHANTABILITY or

  .012384 +.047
::: GP :::
~/Git/llama2.c ➤ make run
gcc -O3 -o run run.c -lm
ld: warning: duplicate -rpath '/gp/usr/li
ld: warning: duplicate -rpath '/gp/usr/li
ld: warning: ignoring duplicate libraries
gcc -O3 -o runq runq.c -lm
ld: warning: duplicate -rpath '/gp/usr/li
ld: warning: duplicate -rpath '/gp/usr/li
ld: warning: ignoring duplicate libraries
  .736043 +.035
::: GP :::
~/Git/llama2.c ➤ ./run ./stories15M.bin
Once upon a time, there was a clever litt
 bag of powder. Max was curious about the
Max's owner was upset and told Max to com
o escape from the powder. But when he got
Max's owner was upset and didn't know wha
p all the powder and throw it at the floo
to be more careful around powder.
achieved tok/s: 146.699267
  2" .149229 +.038
```

```
::: GP :::
~/Git/llama2.c ➤ gcc --version
gcc (Gentoo 13.2.0 p2) 13.2.0
Copyright (C) 2023 Free Software Foundatio
This is free software; see the source for
warranty; not even for MERCHANTABILITY or

  .013717 +.053
::: GP :::
~/Git/llama2.c ➤ make runfast
gcc -Ofast -o run run.c -lm
ld: warning: duplicate -rpath '/gp/usr/lib
ld: warning: duplicate -rpath '/gp/usr/lib
ld: warning: ignoring duplicate libraries
gcc -Ofast -o runq runq.c -lm
ld: warning: duplicate -rpath '/gp/usr/lib
ld: warning: duplicate -rpath '/gp/usr/lib
ld: warning: ignoring duplicate libraries
  .748195 +.037
::: GP :::
~/Git/llama2.c ➤ ./run ./stories15M.bin
Once upon a time, there was a clumsy dog n
 on the ground and he wanted to eat it. Bu
The strawberry rolled away and Spot chased
 But then, something unexpected happened.
Spot's owner laughed and helped him out of
oved to play fetch. Spot's owner gave him
achieved tok/s: 461.538462
  1" .001685 +.038
```

可能的 Baby Llama 2 性能优化点是哪些?

——编译篇之三 · 比较 3

RISCV 下（QEMU）

```
ryan@vm2rv ~/llama2.c $ make run
gcc --version
gcc (Gentoo 13.2.1_p20230826 p7) 13.2.1 2023
Copyright (C) 2023 Free Software Foundation,
This is free software; see the source for co
warranty; not even for MERCHANTABILITY or FI

gcc -O3 -o run run.c -lm
gcc -O3 -o runq runq.c -lm
ryan@vm2rv ~/llama2.c $ ./run stories15M.bin
Once upon a time, there was a little girl na
ommy, it's a sunset!" said Lily. "Yes, it's
As they walked, they saw a busy squirrel. "H
asked Lily. "I think we know what happens wh
Suddenly, Lily and her mommy heard a loud no
 mommy. They ran back to the car and drove h
 the flowers made a mark!" said Lily. "Yes,
achieved tok/s: 3.045831
```

```
ryan@vm2rv ~/llama2.c $ make runfast
gcc --version
gcc (Gentoo 13.2.1_p20230826 p7) 13.2.1 20230
Copyright (C) 2023 Free Software Foundation,
This is free software; see the source for cop
warranty; not even for MERCHANTABILITY or FIT

gcc -Ofast -o run run.c -lm
gcc -Ofast -o runq runq.c -lm
ryan@vm2rv ~/llama2.c $ ./run stories15M.bin
One day, a little boy named Tim went to the p
en to the birds sing.
At the park, Tim saw a big tree with a lot of
he tree." Tim nodded and picked an apple.
As Tim ate the apple, he felt very hot from t
glass of water to drink. Tim drank his water
achieved tok/s: 3.040701
```

```
ryan@vm2rv ~/llama2.c $ make run CC=clang
clang --version
clang version 17.0.6
Target: riscv64-unknown-linux-gnu
Thread model: posix
InstalledDir: /usr/lib/llvm/17/bin
Configuration file: /etc/clang/riscv64-unknown
clang -O3 -o run run.c -lm
clang -O3 -o runq runq.c -lm
ryan@vm2rv ~/llama2.c $ ./run stories15M.bin
Once upon a time, there was a big, dependable
d, a little girl named Lily.
At the park, they found a shiny piece of coppe
o play with. They found a stick and some stick
As the sun went down, it was time for Buddy to
ed his tail. They went home, tired but happy
achieved tok/s: 3.052987
```

```
ryan@vm2rv ~/llama2.c $ make runfast CC=clang
clang --version
clang version 17.0.6
Target: riscv64-unknown-linux-gnu
Thread model: posix
InstalledDir: /usr/lib/llvm/17/bin
Configuration file: /etc/clang/riscv64-unknow
clang -Ofast -o run run.c -lm
clang -Ofast -o runq runq.c -lm
ryan@vm2rv ~/llama2.c $ ./run stories15M.bin
Once upon a time, there was a little girl nam
s of colors. Lily was so happy and wanted to
She tried to match the colors of her dolls, b
 but also excited because she had never seen
She opened the box and found a toy monster in
happy she found a new toy to play with.
achieved tok/s: 3.067540
```

可能的 Baby Llama 2 性能优化点是哪些?

——其他

- 是否有其他可以优化的方面呢?
- 精确度的判断? （自己怎么测）
- 输出位置?
- 当前的内存占用率极少
- 当前的 CPU 利用率极少

有关文本转语音

- 在有限的时间里，我并没有找到现成的轻量级 TTS 方案。
- 只能留给大家思考了

感谢观看

RVSPOC 组委会 – Ryan

2024 年 1 月 2 日