WHY NOT RUN OPENCL-ACCELERATED LLM ON YOUR PHONE?

DAVID HEIDELBERG

WHO AM I?

- Mesa3D developer
  (mostly CI for GPUs testing @Collabora)
- In free time:
  - Mobile/Embedded Linux developer
  - Alpine/Debian/postmarketOS contributor

CONTENT

- Mesa 3D, OpenCL, and framework introduction
- Computation before OpenCL, ROCm, CUDA,...
- What we have now
- How it works (on the phone)
- Where compute heading?

BASIC TERMS

OpenCL (Open Computing Language) is a framework for writing programs that execute across heterogeneous platforms consisting of central processing units (CPUs), graphics processing units (GPUs), digital signal processors (DSPs)....

Mesa 3D is open source implementations of OpenGL, OpenGL ES, Vulkan, OpenCL, and more!

tinygrad is autograd engine with backpropagation (tinygrad can run LLaMA and Stable Diffusion)

COMPUTATION BEFORE "COMPUTE"

- Slow: CPU (branching, predictions, complexity)
- Hard: OpenGL was used to do parallel calculations

WHAT HARDWARE IS AVAILABLE TODAY

- CPU: complex and slow
- GPU: simpler and reasonable fast
- NPU: specialized and fast or power efficient

WHAT TYPE OF SOFTWARE USES COMPUTE

- Image processing
- Scientific software
- Hash calculations (currently there are already DSPs for that), password cracking
- Large Language Models

THE DRIVERS

- 2012: Clover implmentation
- 2022: The RustICL OpenCL Library is conformant with OpenCL 3.0.
- Supported: Intel (iris), AMD (r600, radeonsi), Mali (panfrost), on Vulkan (zink), CPU (llvmpipe)
- Work in progress: Apple (Asahi), **Qualcomm Adreno** (freedreno), Vivante (etnaviv), RaspberryPi (v3d)

TINYGRAD ON A LINUX PHONE

```
$ export RUSTICL_ENABLE=msm && export GPU=1 && export PYTHONPATH="."
$ python examples/gpt2.py --model_size=gpt2 --prompt "Hello, how are you?" --count 11 --temperature 0
using GPU backend
using gpt2
ram used:  0.50 GB, lm_head.weight                                    : 100%|█| 149/149 [00:00<
loaded weights in 360.81 ms, 0.65 GB loaded at 1.81 GB/s
100%|██████████████████████████████████████████████████████████| 11/11 [00:11<00:00,  1.05s/it]
Generating text...
Hello, how are you?

I'm a little bit of a nerd.
```

TINYGRAD BENCHMARK

OnePlus 6, Snapdragon 845.

- CPU: Octa-core Kryo 385 (4x2.8 GHz & 4x1.7 GHz)
- GPU: Adreno 630 (@710 Mhz)

option	time (ms)	CPU load
GPU=0	1210.89	8x 100%
GPU=1	390.45	2x +- 20%

HOW TO MAKE IT WORK (POSTMARKETOS)

1. Install `mesa-rusticl` from Alpine MR (Mesa3D patched for freedreno support)
2. Using tool like `mrtest` it's easy to install the artifacts.
3. Download tinygrad.
4. Run with the right arguments.

WHERE COMPUTE IS HEADING?

- OpenCL getting widely supported on Linux.
- Has to catch up with CUDA (watch Dave Airlie talk)
- Clover implementation is dead
- Even low-power device as a phone can offer useful compute.

CREDITS

Karol Herbst for bringing rusticl alive!

Rob Clark for working on freedreno!

Dmitry Baryshkov for preparing freedreno for rusticl integration!

... and many others who contributed to Mesa 3D development.

THAT'S ALL FOLKS!

Thank you for your attention! Questions?

https://floss.social/@okias
https://x.com/okiasx
https://www.linkedin.com/in/okias/