RK3588 TTS

#whoami - Martin

- 本職： C++ / HPC / Systems software
- Open source dev
- Do what people do

- C++ web framework maintainer - drogonframework/drogon
- Patches OSS for OpenBSD
- AI work at times

GitHub: https://github.com/marty1885

Blog: https://clehaxze.tw/

Disclaimer

- The opinions expressed in these slides are personal and do not represent the views or opinions of my employer. Any references to specific products, services, or organizations are for illustrative purposes only and do not imply endorsement. Please consult with appropriate professionals or your organization for specific advice related to your circumstances.

RK3588

- Rockchip’s flagship SoC
- 8 Cores (4 BIG + 4 little)
- 3 NPU cores (Neural processor)
- 3rd party Ubuntu 22.04 port
- Low power, 7W TDP

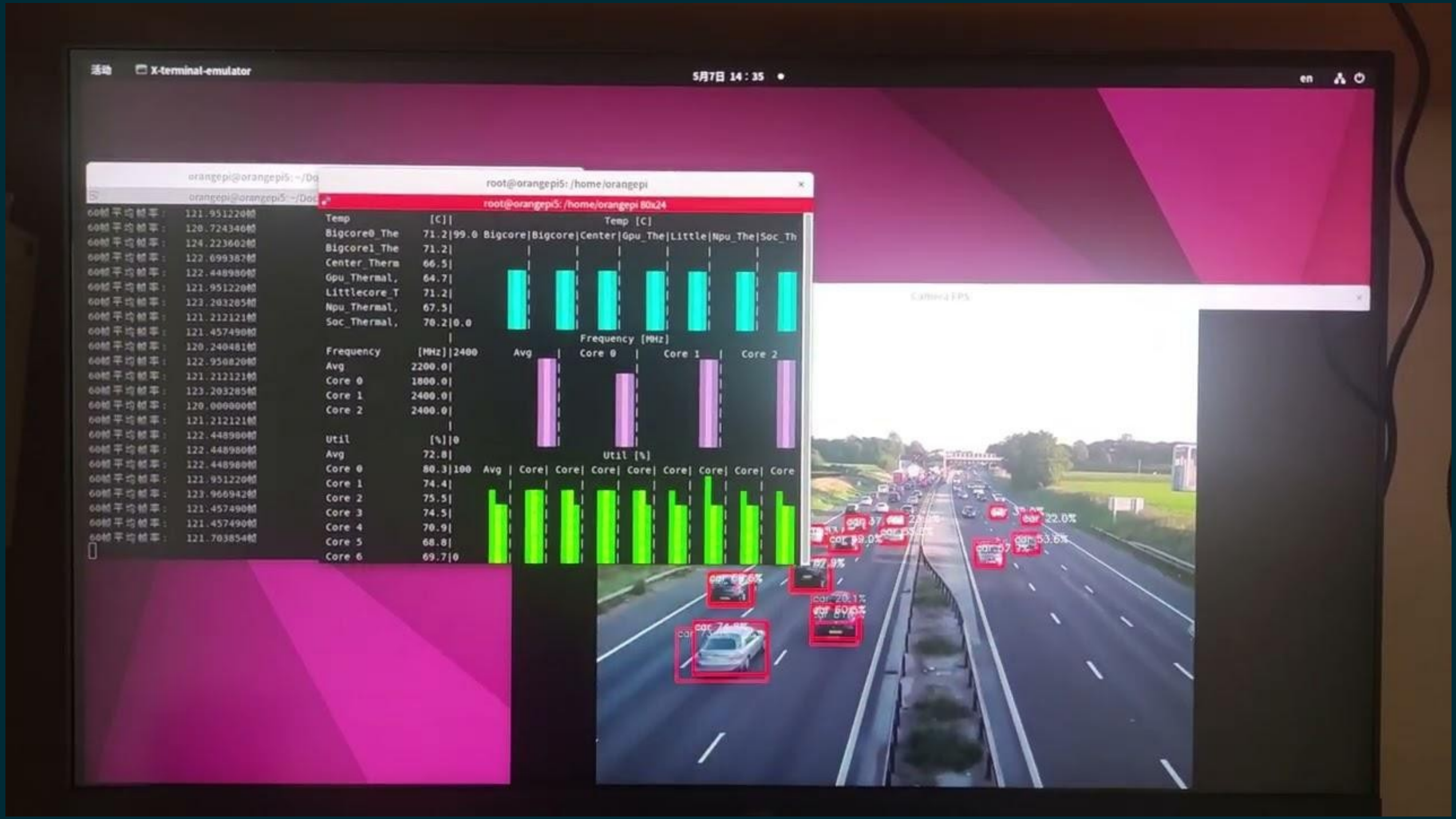
- Closed SDK :(

- Bought for porting LLM, but another story sometime

RK3588 (cont..)

- NPU Designed to run YOLO or ResNet
- Not designed for speech synthesis

Credit: 李安 under fair use. YouTube VID: wwRSP9ucbhw



orangepi@orangepi5: ~/Do

```
60帧平均帧率：   121.951220帧
60帧平均帧率：   120.724346帧
60帧平均帧率：   124.223602帧
60帧平均帧率：   122.699387帧
60帧平均帧率：   122.448980帧
60帧平均帧率：   121.951220帧
60帧平均帧率：   123.203285帧
60帧平均帧率：   121.212121帧
60帧平均帧率：   121.457490帧
60帧平均帧率：   120.240481帧
60帧平均帧率：   122.950820帧
60帧平均帧率：   121.212121帧
60帧平均帧率：   123.203285帧
60帧平均帧率：   120.000000帧
60帧平均帧率：   121.212121帧
60帧平均帧率：   122.448980帧
60帧平均帧率：   122.448980帧
60帧平均帧率：   122.448980帧
60帧平均帧率：   121.951220帧
60帧平均帧率：   123.966942帧
60帧平均帧率：   121.457490帧
60帧平均帧率：   121.457490帧
60帧平均帧率：   121.703854帧
```

root@orangepi5: /home/orangepi

root@orangepi5: /home/orangepi 80x24

```
Temp              [C]|                       Temp [C]
Bigcore0_The     71.2|99.0 Bigcore|Bigcore|Center|Gpu_The|Little|Npu_The|Soc_Th
Bigcore1_The     71.2|
Center_Therm     66.5|
Gpu_Thermal,     64.7|
Littlecore_T     71.2|
Npu_Thermal,     67.5|
Soc_Thermal,     70.2|0.0
                     |                 Frequency [MHz]
Frequency       [MHz]|2400     Avg   |  Core 0   |   Core 1  |   Core 2
Avg            2200.0|
Core 0         1800.0|
Core 1         2400.0|
Core 2         2400.0|
                     |
Util              [%]|0
Avg              72.8|                   Util [%]
Core 0           80.3|100  Avg | Core| Core| Core| Core| Core| Core| Core| Core
Core 1           74.4|
Core 2           75.5|
Core 3           74.5|
Core 4           70.9|
Core 5           68.8|
Core 6           69.7|0
```

Motivation

- Goal : Component of my own digital assistant
- Issue : GPUs are too hot.
    - I can’t handle a 300W heat source in my room in summer

Piper

- https://github.com/rhasspy/piper
- High quality, fast TTS
- CPU: ~5.5x realtime
- GPU: ~40~100x realtime

- RK3588 CPU: 1.1x realtimes

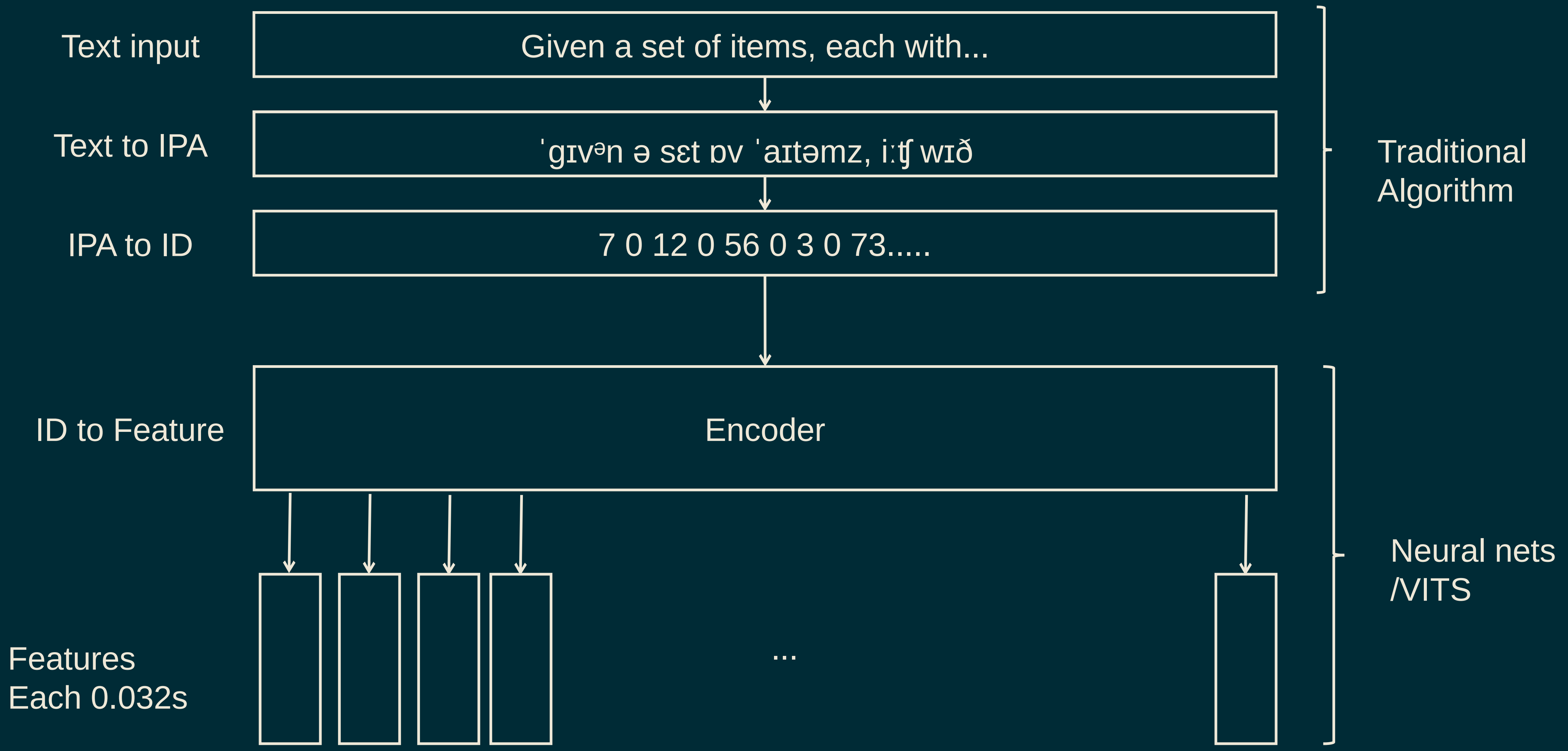
- More important : Piper runs per-sentence. 0.9s delay on 1s sentence
- Horrible UX

Can we do better?

- “ONNX streaming support” - piper #255
- Disects Piper into 2 parts
- Encoder & decoder
- Encoder still can’t run on the NPU
- But.. Decoder can
- And Decoder takes the majority of time!
- Synthesises chunked at 0.032s!!

- Solves the high latency

Piper streaming archicture



Text input

Given a set of items, each with...

Text to IPA

ˈɡɪvᵊn ə sɛt ɒv ˈaɪtəmz, iːʧ wɪð

IPA to ID

7 0 12 0 56 0 3 0 73.....

Traditional Algorithm

ID to Feature

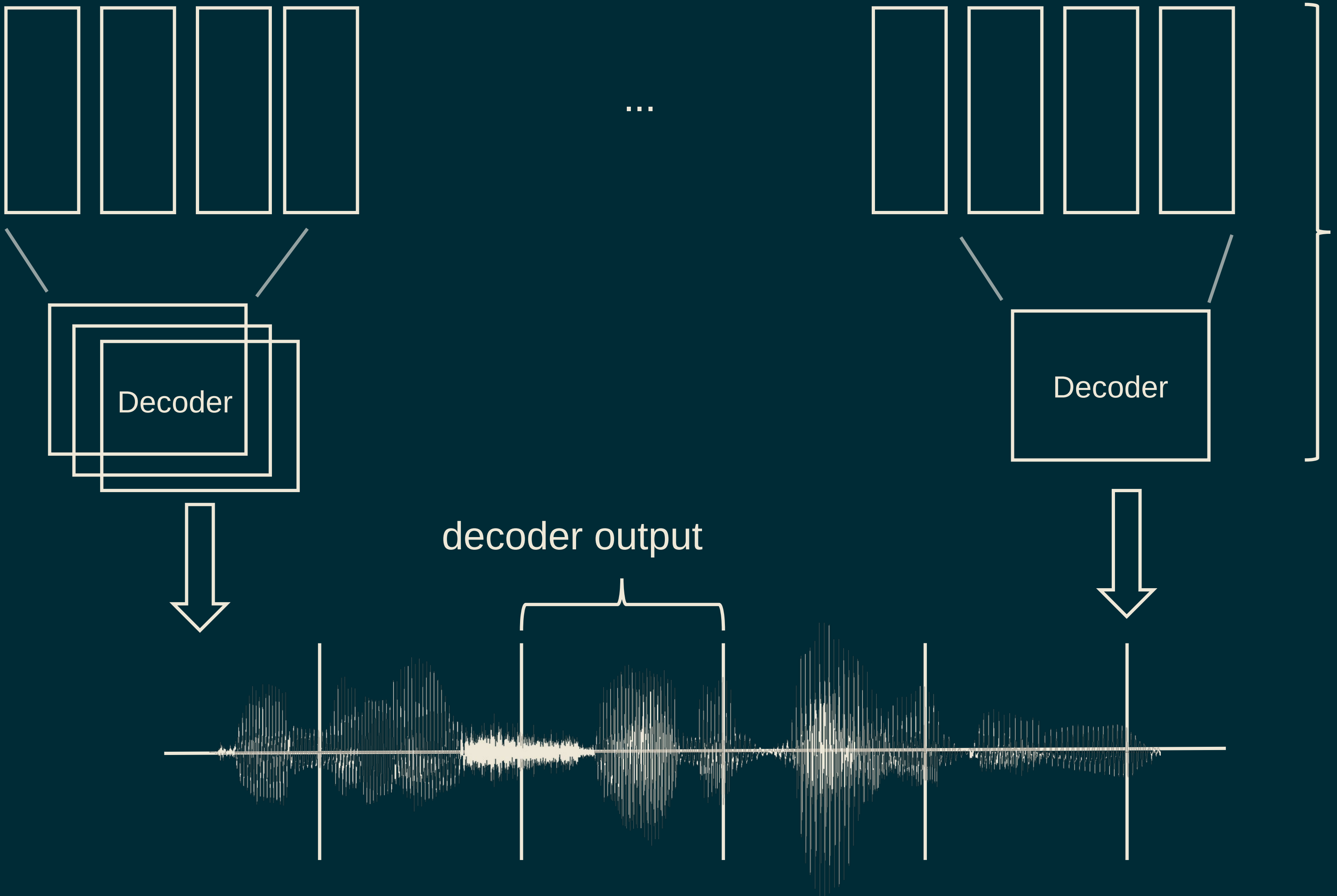
Encoder

Features
Each 0.032s

...

Neural nets /VITS

Piper streaming archicture (cont.)



Features
Each 0.032s

Grouped features
Decoder synthesises
to audio

Combined into audio
stream

Neural nets
/VITS

We can do better

- Req for acceleration
  - The thing to accelerate is slow
  - The slow stuff takes forever
- decoder looks like a vision model
  - 2D matrix input (WxH / features x N sets)
  - Output list of number(class prob / audio)
  - Mostly Convolution

- Decoder is slow
- Decoder looks like vision models

If it walks like a duck and it quacks like a duck

- Then it is a duck

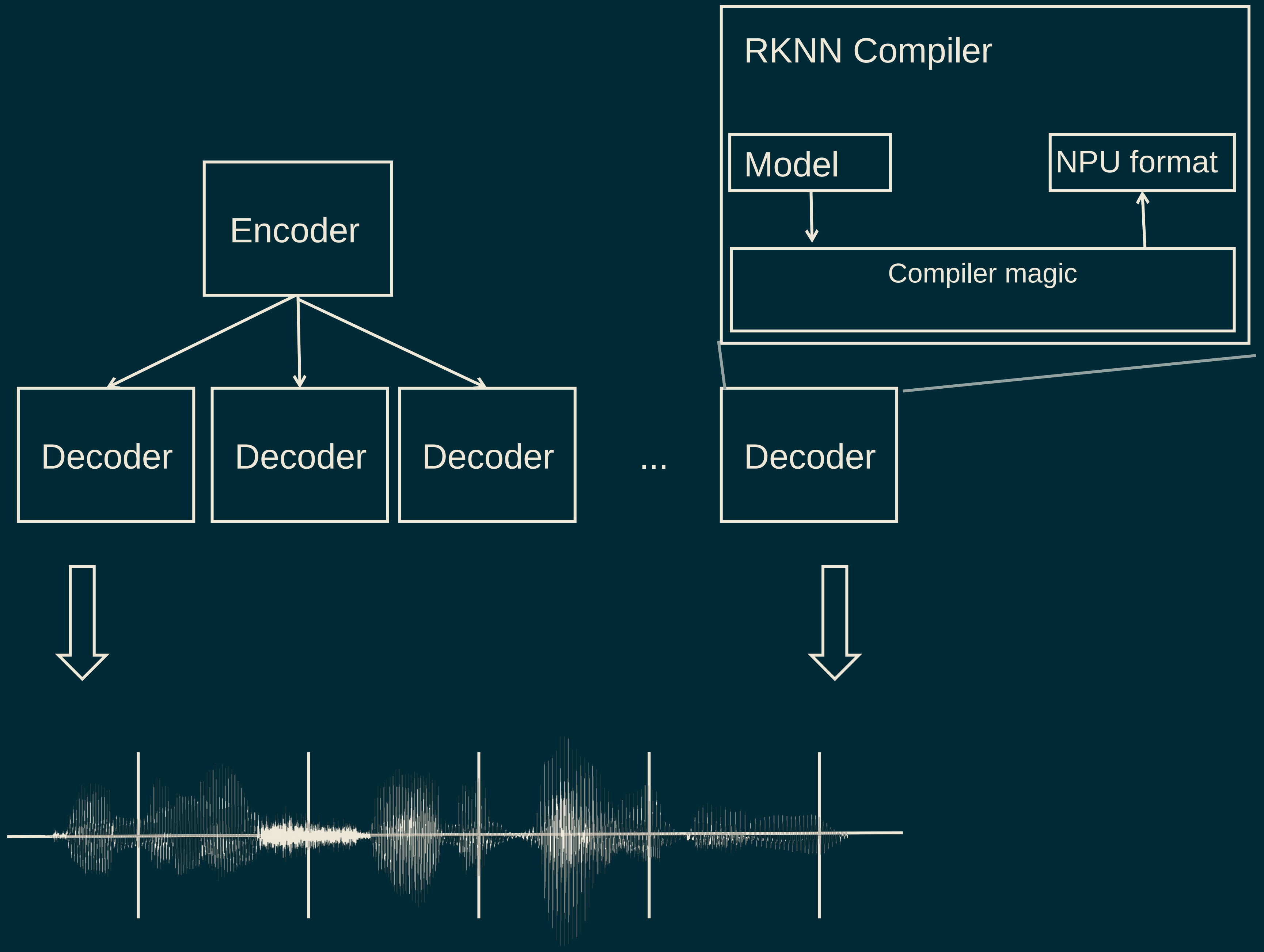
- It is close enough it’ll work
- Why not try?

Much hacking later

Can’t do.
Keep on CPU

Run them on
NPU

Same audio



RKNN Compiler

Model

NPU format

Compiler magic

Encoder

Decoder

Decoder

Decoder

...

Decoder

Done

- https://github.com/marty1885/paroli
- NPU runs at 9x realtime
- Faster then desktop CPU!!!!

- Tech improves UX of AI applications
- Low power enables use cases prev impossible (home, etc..)
- Only works because open source

Future work

- AMD released XDNA driver today
- Intel 14th Gen has NPU

Thank you

- More to come..