From NCW23 to production

NORDUnet Speech2text

NORDUnet
Nordic Gateway for Research & Education

Markus Krogh

NCW23

- Media Track talk about Whisper AI subtitle service at Oslo University
- Whisper.cpp had just come out
- Trying it out in the hallway

```
WEBVTT

00:00:00.000 --> 00:00:08.000
 Mit navn er Markus Krog, og vi står her til NCW 2223 og prøver at lege med Whisper.

00:00:08.000 --> 00:00:10.000
 Nu må vi se, hvordan det går på dansk.

00:00:10.000 --> 00:00:17.000
 Der var noget med, at der skulle være en god mængde data, så lad os se, om vi kan komme op på et halvt sekund.

00:00:17.000 --> 00:00:19.000
 - Et halvt sekund? - Ja, et halvt sekund.

00:00:19.000 --> 00:00:21.000
 Eller så i hvert fald et halvt minut.

00:00:21.000 --> 00:00:24.000
 - Er det bedre? - Nu finder vi ud af, hvordan den klarer det her.

00:00:24.000 --> 00:00:27.000
 Især når der er flere, der snakker. Ja, det tror jeg ikke.
```

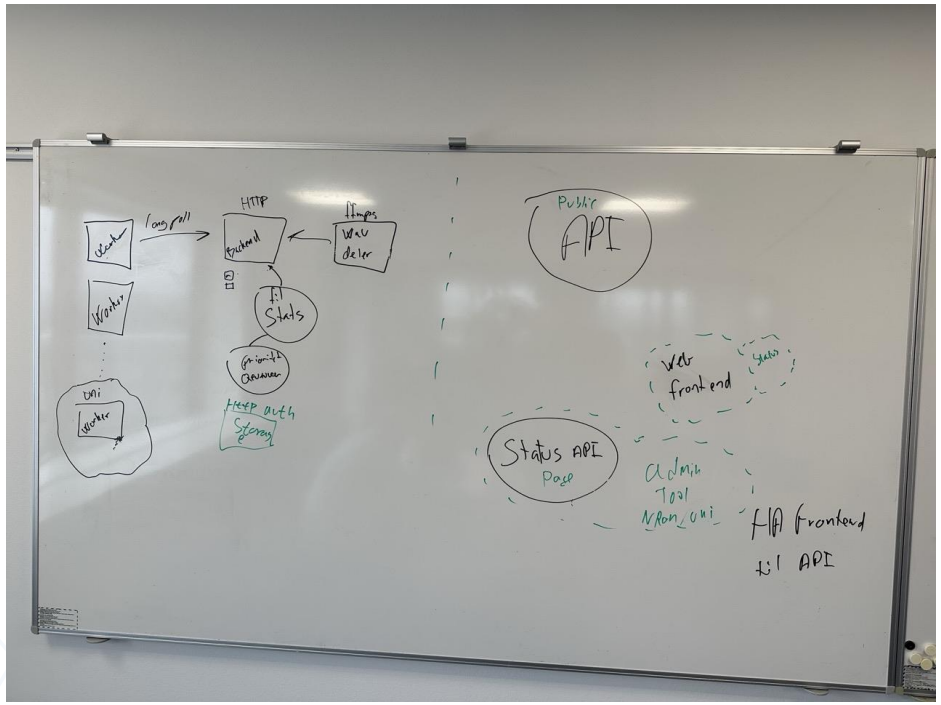
NCW23

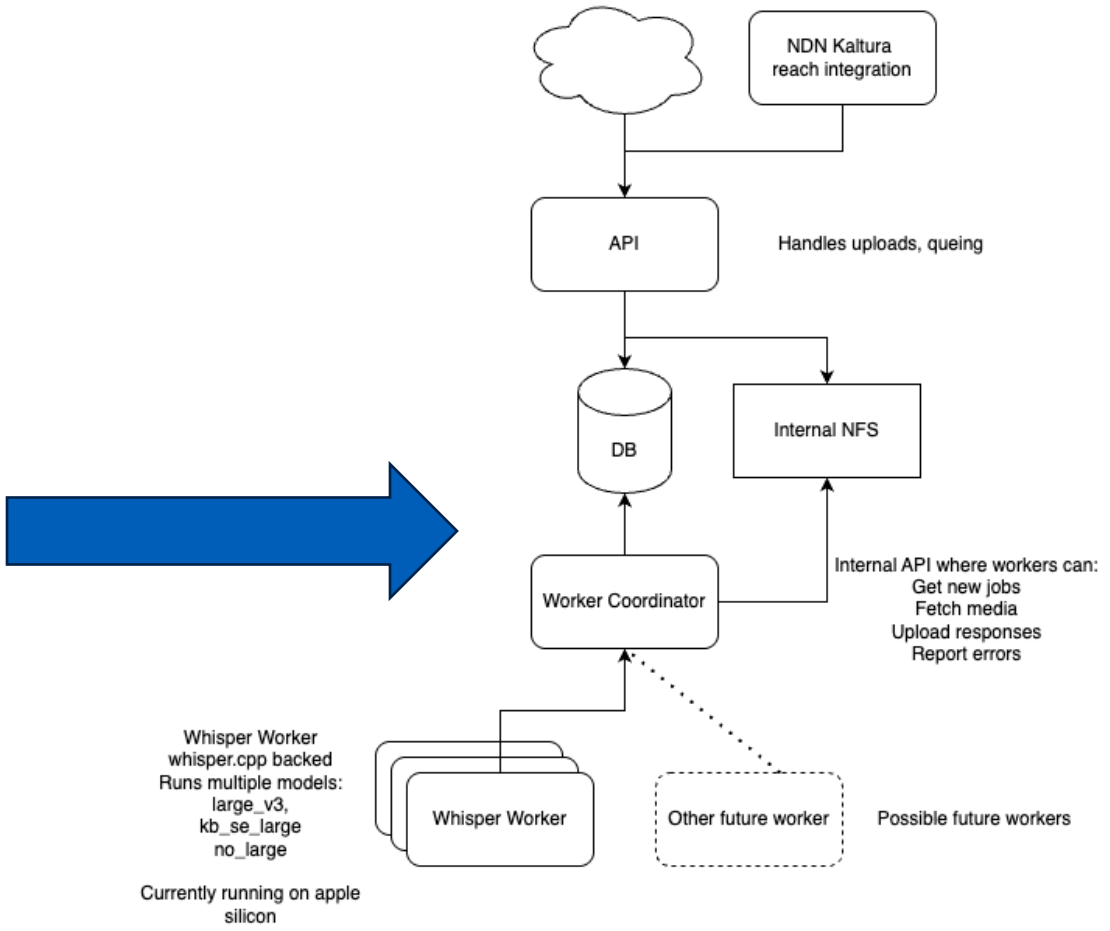
- Media Track talk about Whisper AI subtitle service at Oslo University
- Whisper.cpp had just come out
- Trying it out in the hallway

```
WEBVTT

00:00:00.000 --> 00:00:08.000
 Mit navn er Markus Krog, og vi står her til NCW 2223 og prøver at lege med Whisper.

00:00:08.000 --> 00:00:10.000
 Nu må vi se, hvordan det går på dansk.

00:00:10.000 --> 00:00:17.000
 Der var noget med, at der skulle være en god mængde data, så lad os se, om vi kan komme op på et halvt sekund.

00:00:17.000 --> 00:00:19.000
 - Et halvt sekund? - Ja, et halvt sekund.

00:00:19.000 --> 00:00:21.000
 Eller så i hvert fald et halvt minut.

00:00:21.000 --> 00:00:24.000
 - Er det bedre? - Nu finder vi ud af, hvordan den klarer det her.

00:00:24.000 --> 00:00:27.000
 Især når der er flere, der snakker. Ja, det tror jeg ikke.
```

Sketching out a design

- Multiple services
  - Seperate workers
- Simple API
- Allow workers to run anywhere – local university deploy





Speech2Text API v01 OAS 3.0

http://10.96.10.50/swagger/v1/swagger.json

NORDUnet Speech2Text API

Terms of service
NORDUnet - Website

Authorize

Get

- GET /api/s2t/tasks Get allResults tasks for a billing reference
- GET /api/s2t/status/{uuid} Get task from Id

Post

- POST /api/s2t/status/{uuid} Update task Status
- POST /api/s2t/task/{uuid} Setting Task info
- POST /api/s2t/upload/{uuid} Sending a file

Put

- PUT /api/s2t/task Add new task

Speech2TextAPI

- GET /

Subtiles for days

- Testing large_v3 on 3m6s audio
  - **EPYC zen3 64 treads machine**
    - 1 instance 6m49s ~x0.5
    - With 8 threads ~x0.66
  - **Zen4 desktop + 4070 rtx**
    - CPU only
      - 1 instance 2m26s ~x1.2
      - 2 instances 2m41s ~x1.1
      - 3 instances 4m10s ~x0.7
    - CUBLAS (~422mb vram)
      - 1 instance 56s ~x3.3
      - 2 instances 1m27s ~x2.1
  - **M2 laptop**
    - 1 instance 46s ~x4 (was 26s ~x7.8)
    - 1h20m audio ~x9.8
  - **MacStudio (1h20m audio)**
    - 1 instance ~x11-x12
  - **Mac mini (1h20m audio)**
    - 1 instance 5m25s ~x12

Usage

- In production since april 2025 for Kaltura
- Last 30 days