Technische Universität Berlin

Nextcloud AI Erweiterungen – Konzepte gegen den GPU-Hunger

Dr. Thomas Hildmann | Technische Universität Berlin | 80. DFN Betriebstagung – Forum Cloud

Agenda

1. **Einleitung**
2. Schneller **Überblick** über Featureset
3. **GPU as a Sevice** in der Public und Private Cloud
4. Was ist der **Plan**?
5. **Fazit**

Neben Bullshit-Bingo – Wovon reden wir?

- Nextcloud 27 bringt **neue Features** mit, die auf **AI Algorithmen** basieren
- Meine Aussage Nextcloud Advisory Board: „Für mich ist **wichtiger**, **welche Use-Cases** ich mit neuen Features abdecken kann, **als** auf **welcher Algorithemklasse** sie basieren.“ (sinngemäß)
- Spannend ist jedoch die Frage: **Wie skaliert das?**
  - Auf meinem Mac mini in Ubuntu-Nachnutzung sieht man nichts.
  - Aber mit 4.000+ parallelen Sessions kann das anders aussehen
- Was machen wir also **architekturell**? Und **wie** können wir evtl. **mit** der **Last umgehen**, wenn sie denn signifikant wird?

Ethical AI Rating

Drei Bedingungen

1. Ist die Software Open Source (Nutzung und Training)
2. Gibt es ein freies Modell zum Selbst-Hosting?
3. Sind die Trainingsdaten frei und transparent?

Vier Wertungen

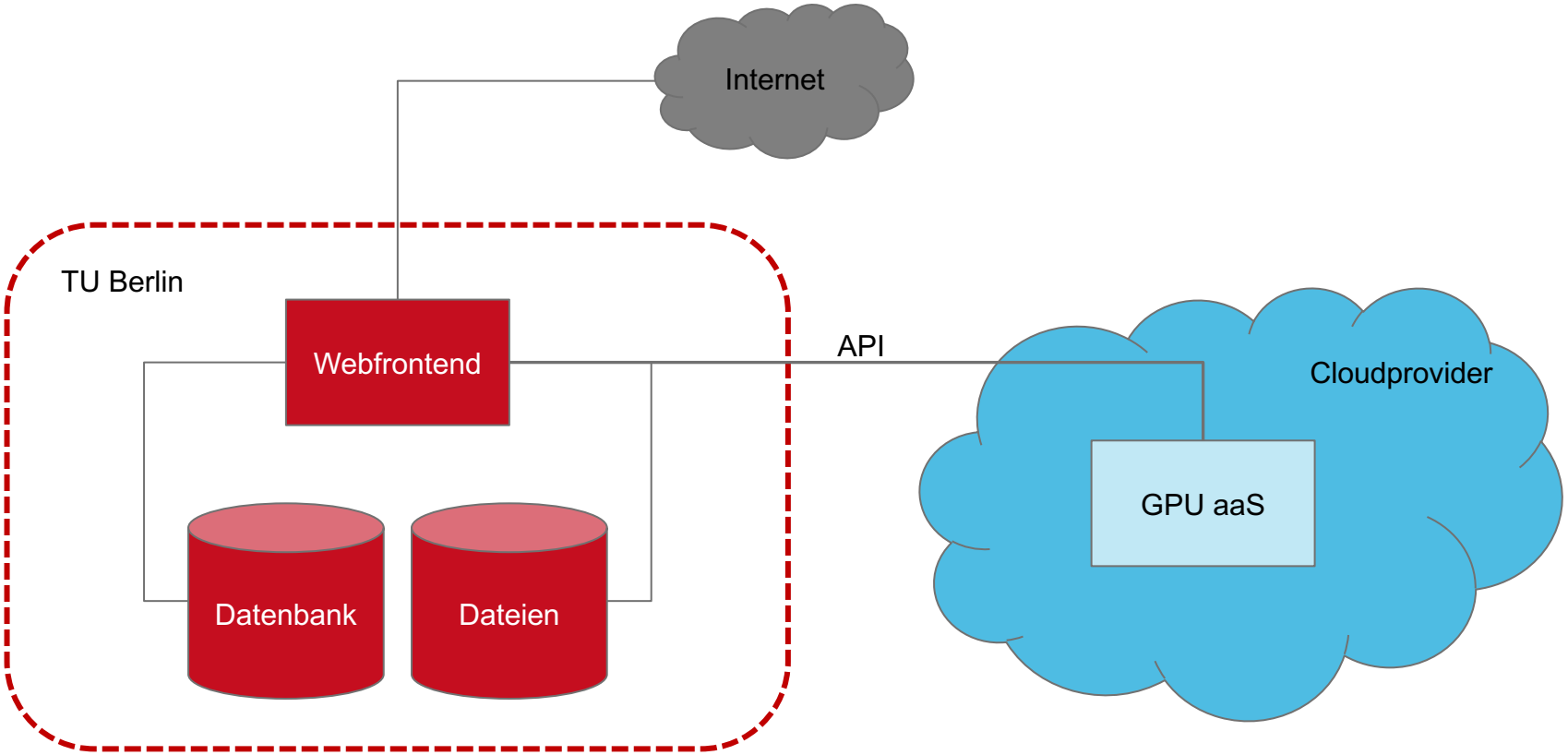
1. Grün: Alle Bedingungen erfüllt.
2. Gelb: Zwei Bedingungen erfüllt.
3. Orange: Eine Bedingung erfüllt.
4. Rot: Keine Bedingung erfüllt.

AI Features beginnend mit Nextcloud 27

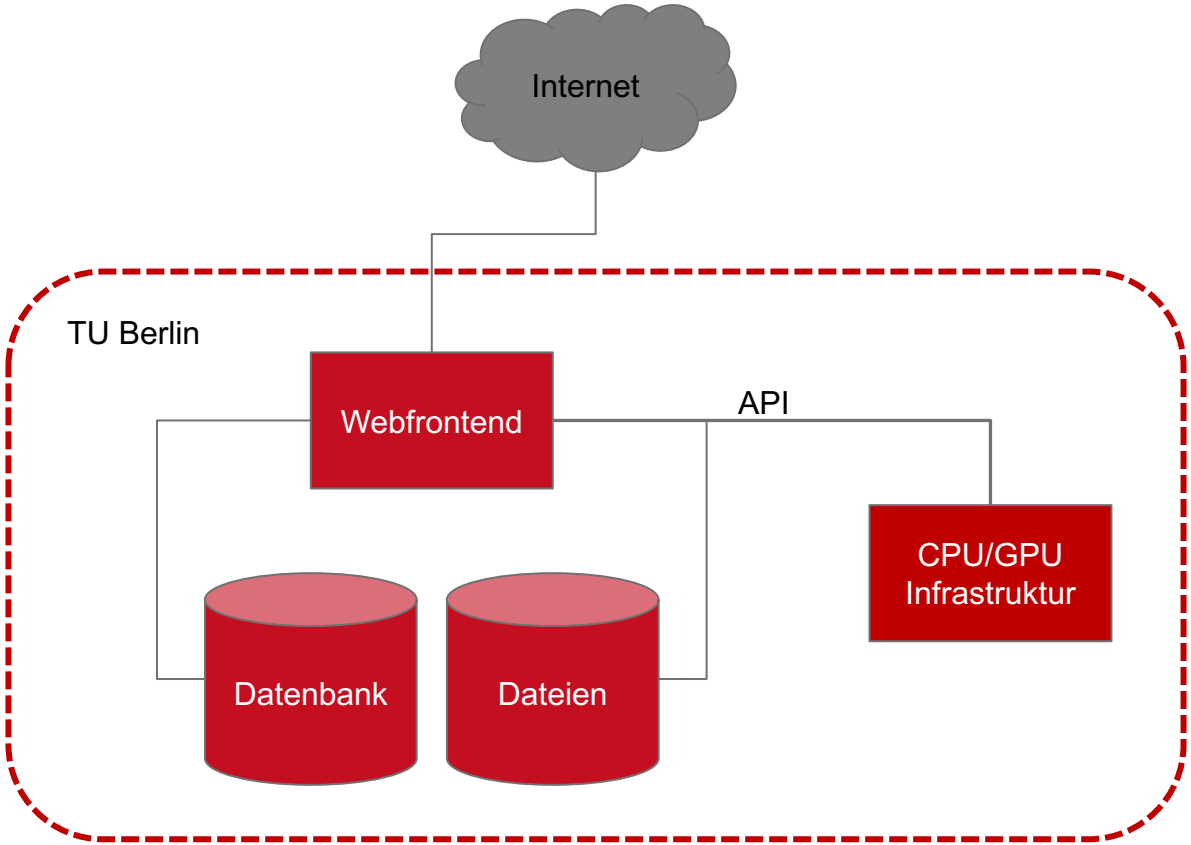
Smart inbox	Mail	Green
Image object recognition	Recognize	Green
Image face recognition	Recognize	Green
Video action recognition	Recognize	Green
Audio music genre recognition	Recognize	Green
Suspicious login detection	Suspicious Login	Green
Related resources	Related Resources	Green
Recommended files	recommended_files	Green
Machine translation	Translate	Green
	LibreTranslate integration	Green
	DeepL integration	Red
	OpenAI and LocalAI integration (via OpenAI API)	Red
	OpenAI and LocalAI integration (via LocalAI)	Green
Speech-To-Text	Whisper Speech-To-Text	Yellow
	OpenAI and LocalAI integration	Yellow
	Replicate integration	Yellow

Image generation	Local Stable Diffusion	Yellow
	OpenAI and LocalAI integration (via OpenAI API)	Red
	OpenAI and LocalAI integration (via LocalAI)	Yellow
	Replicate integration	Yellow
Text generation	Local large language model (via GPT4all Falcon)	Green
	Local large language model (via Llama 2)	Yellow
	OpenAI and LocalAI integration (via OpenAI API)	Red
	OpenAI and LocalAI integration (via LocalAI)	Green
Context Chat	Nextcloud Assistant Context Chat	Yellow
	Nextcloud Assistant Context Chat (Backend)	Yellow

GPU as a Service



GPU as a Service



Nächste Schritte

1. **Evaluation** im kleinen Kreis:
   - **Features: Use-Cases** an der TU Berlin
2. **PoC** mit Cloud-Provider
3. Tests mit **eigener Infrastruktur**
4. **Auswertung**: Metriken, Nutzungsprofile, Erkenntnisse
5. **Angebot** für DFN-Cloud, evtl. mit eigenem Preisschild

Fazit

- Neue Nextcloud **AI Features sind** erst einmal nur neue **Features**
- Wir **evaluieren** zunächst, **welche Potentiale** darin stecken.
- Bei den **Tests** an der TU Berlin werden gleich **verschiedene Infrastrukturkonzepte** geprüft.
- **Am Ende berichten wir und** machen ein **DFN-Cloud Angebot** daraus.