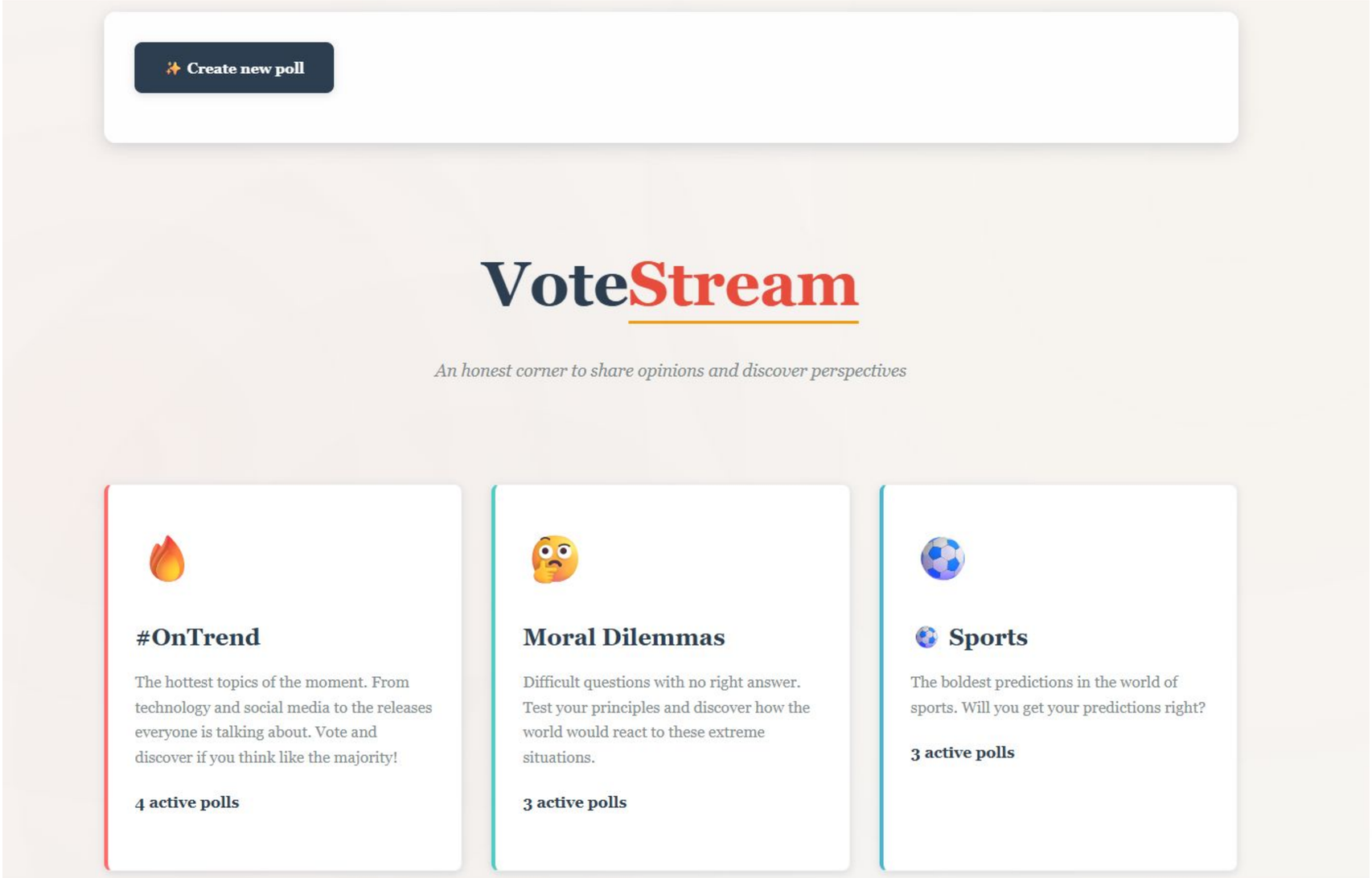
VoteStream: A Scalable Prototyping Project



Technische Universität Berlin

What is VoteStream?

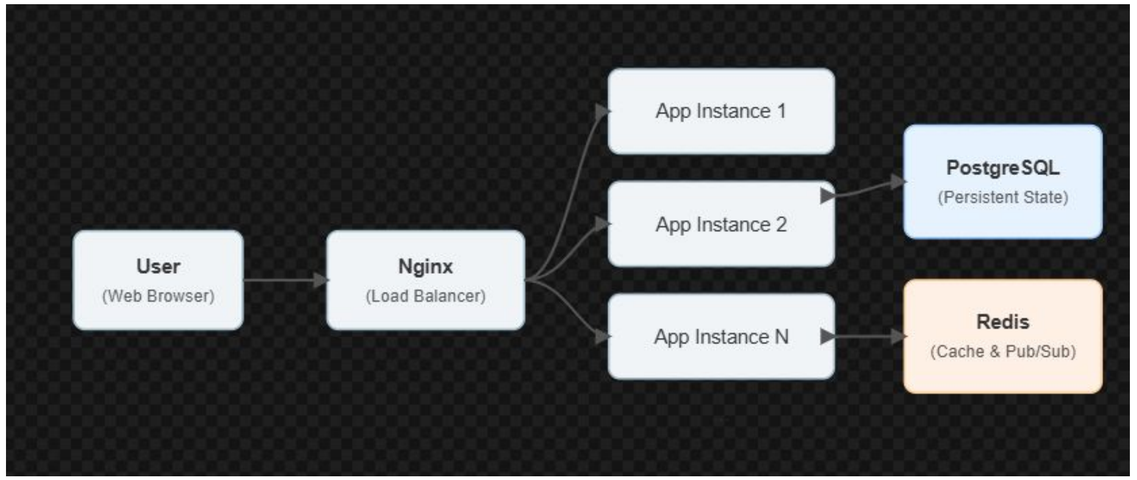
·VoteStream is a real-time, scalable polling application.

·Our motivation was to create a platform where people can genuinely express their opinions and see what the community really thinks about trending topics, moral dilemmas, and more.

·The platform provides instant feedback, with results updating live for everyone as votes are cast.

Technische Universität Berlin

System Architecture



Our system uses a decoupled, stateless architecture for high scalability. The main components are:

- **FastAPI:** The core web framework for the application.
- **PostgreSQL:** Handles the persistent storage of all polls and votes.
- **Redis:** Serves a dual role for caching to speed up responses and for real-time messaging via Pub/Sub.
- **Nginx:** Acts as the reverse proxy and load balancer, distributing traffic to the application instances.

Technische Universität Berlin

Requirement 1: State Management

Requirement: "Your application must manage some kind of state."

Our Solution:

·All application state (polls, options, votes) is durably persisted in a PostgreSQL database.

·We use SQLModel for clear and robust data mapping.

Technische Universität Berlin

Requirement 2: Horizontal & Vertical Scaling

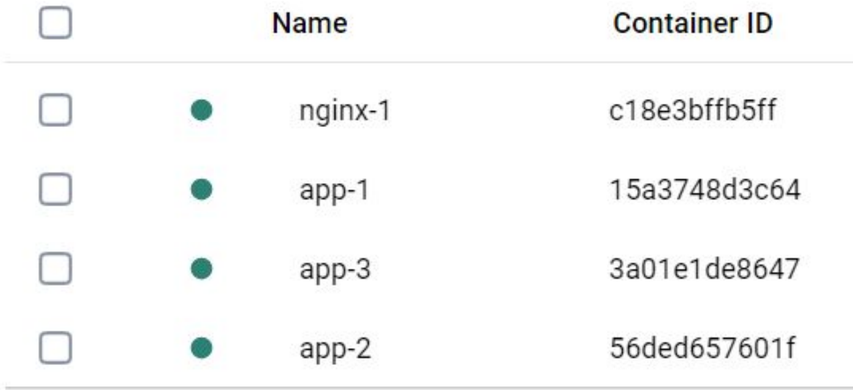
Requirement: "Your application needs to be able to scale vertically and horizontally."

Our Solution:

```
command: >
  sh -c "python migrate_indices.py &&
         python seed_polls.py &&
         gunicorn main:app -w 4 -k uvicorn.workers.UvicornWorker
         --bind 0.0.0.0:8000
         --access-logfile -
         --error-logfile -
         --worker-connections 1000
         --max-requests 1000
         --max-requests-jitter 100"
```

·Vertical Scaling: Achieved by running the app with multiple Gunicorn workers, allowing it to use multiple CPU cores.

·Horizontal Scaling: Enabled by our stateless application design. As you can see, we can run multiple container instances behind the Nginx load balancer.



Technische Universität Berlin

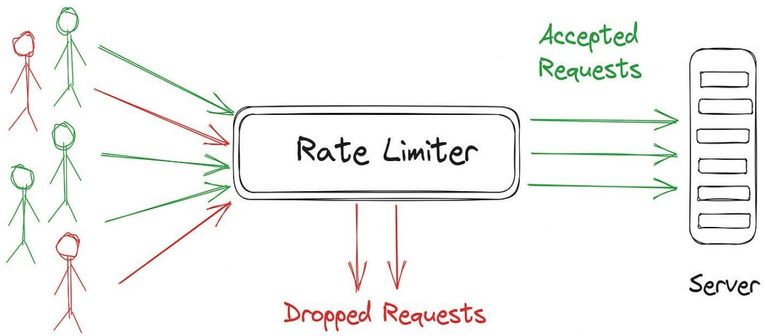
Requirement 3: Overload Mitigation

Requirement: "When your application is scaled up/out, it should not be possible to overload another component."

Our Solution:

·Rate Limiter: A custom middleware prevents abuse by limiting requests per client.

·Circuit Breaker: A custom-built pattern protects our database and cache from cascading failures during high stress or service outages.



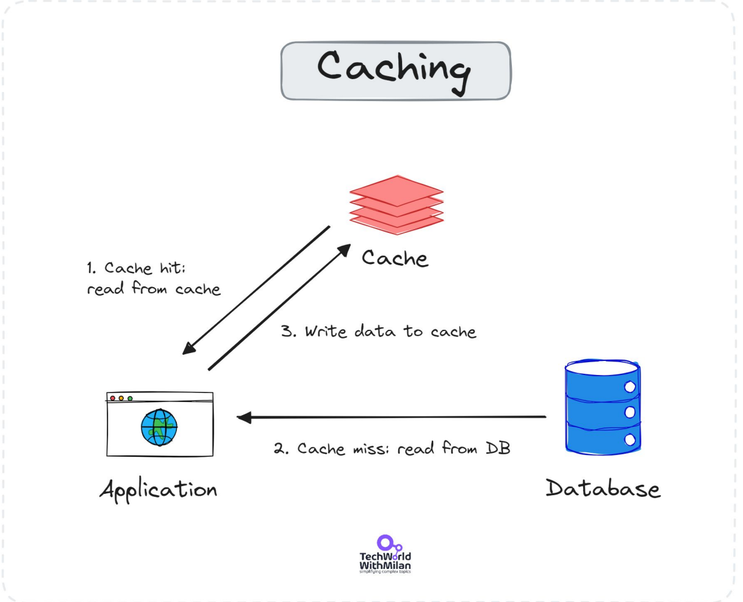
Technische Universität Berlin

Requirement 4: Two Additional Strategies

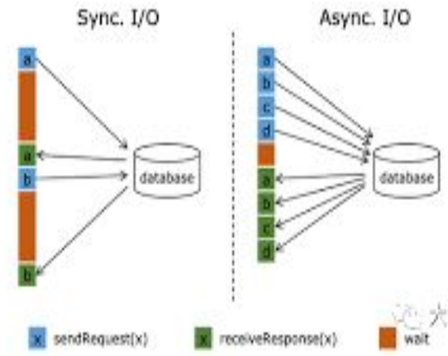
Requirement: "Implement two more strategies...covered either in the lecture content or during the presentations."

Our Solution:

·Caching: We use Redis extensively as a cache-aside layer to reduce database load and serve frequent requests from memory.

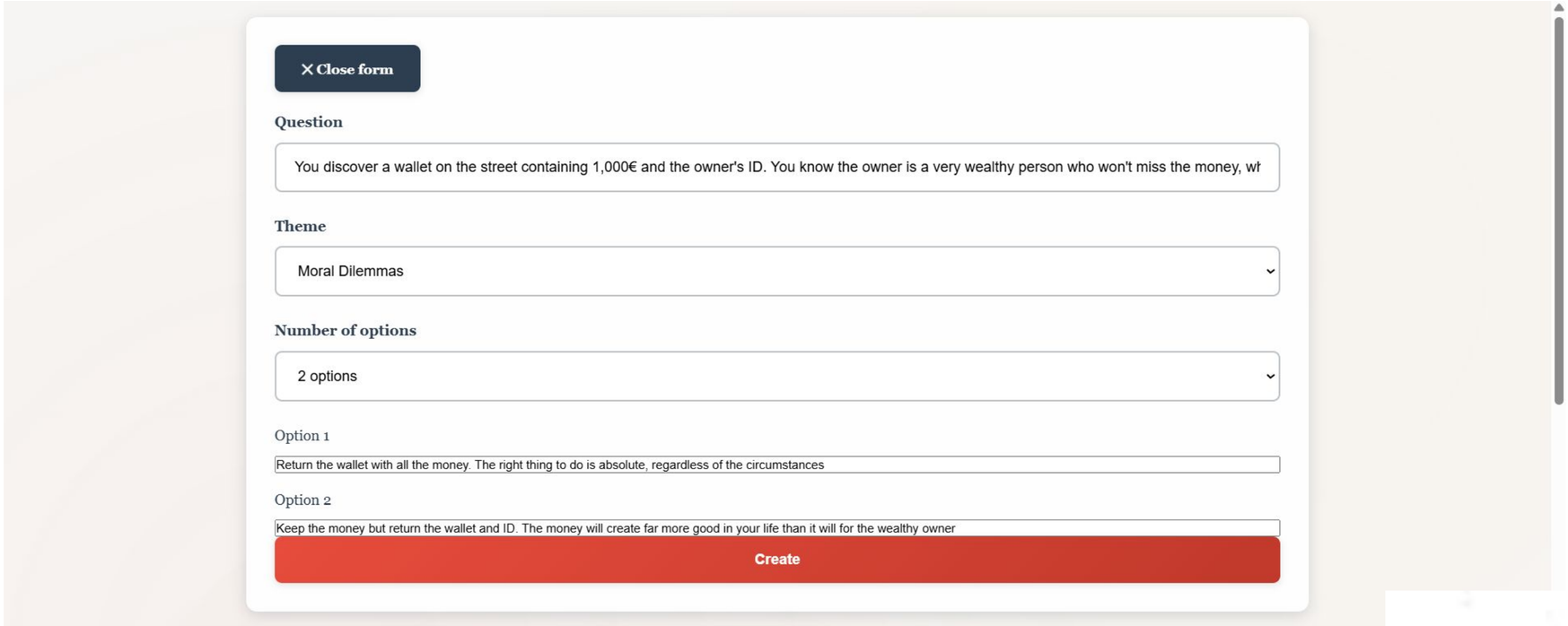


·Asynchronous I/O: The entire backend is built with FastAPI, an async framework that handles thousands of concurrent connections (like WebSockets) efficiently without blocking.



Technische Universität Berlin

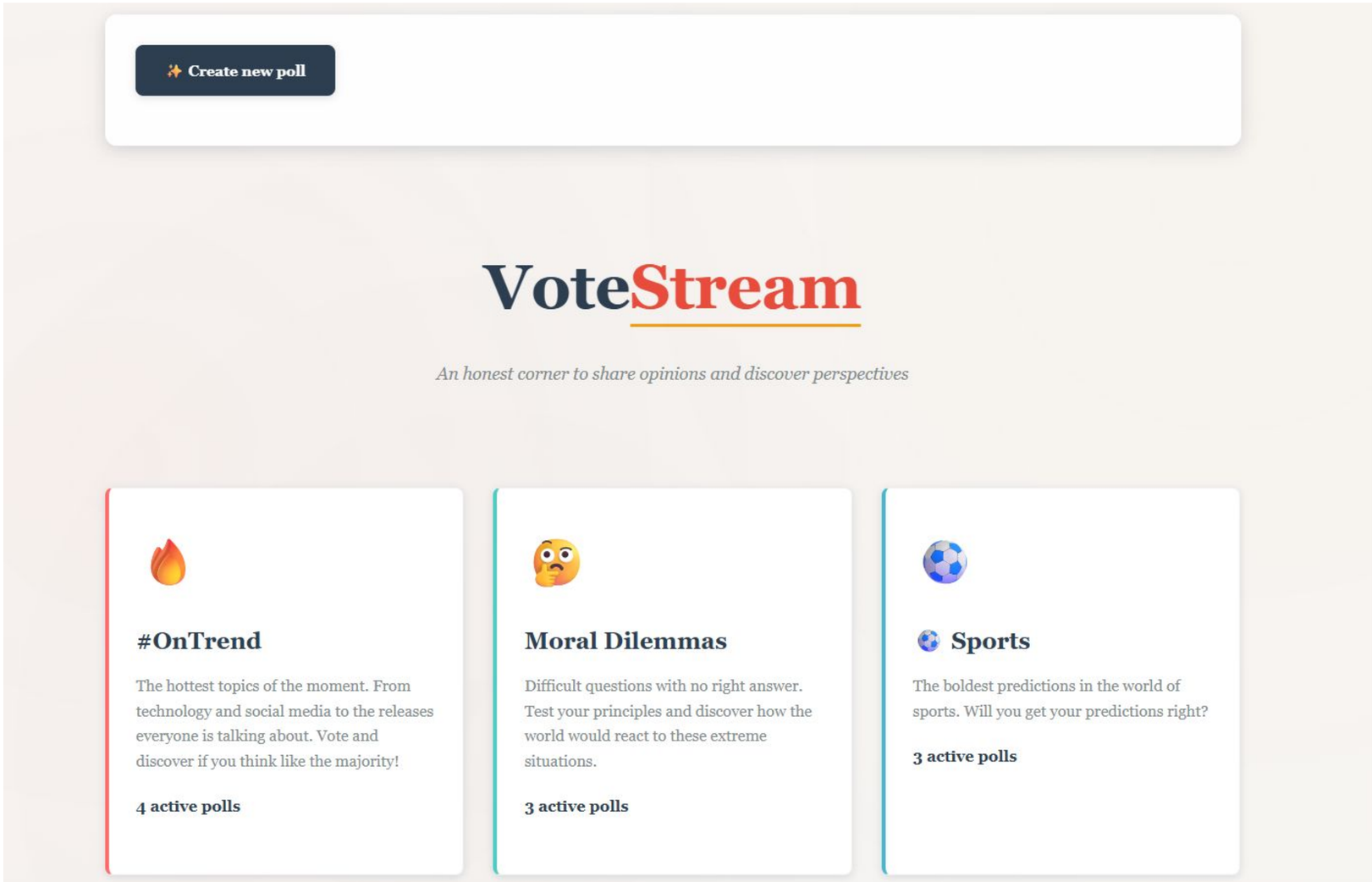
Creating a Poll – State Management



- The user can create a new poll from the frontend.
- This action is sent to the backend, which stores it in the PostgreSQL database using SQLModel.
- This demonstrates persistent state management.

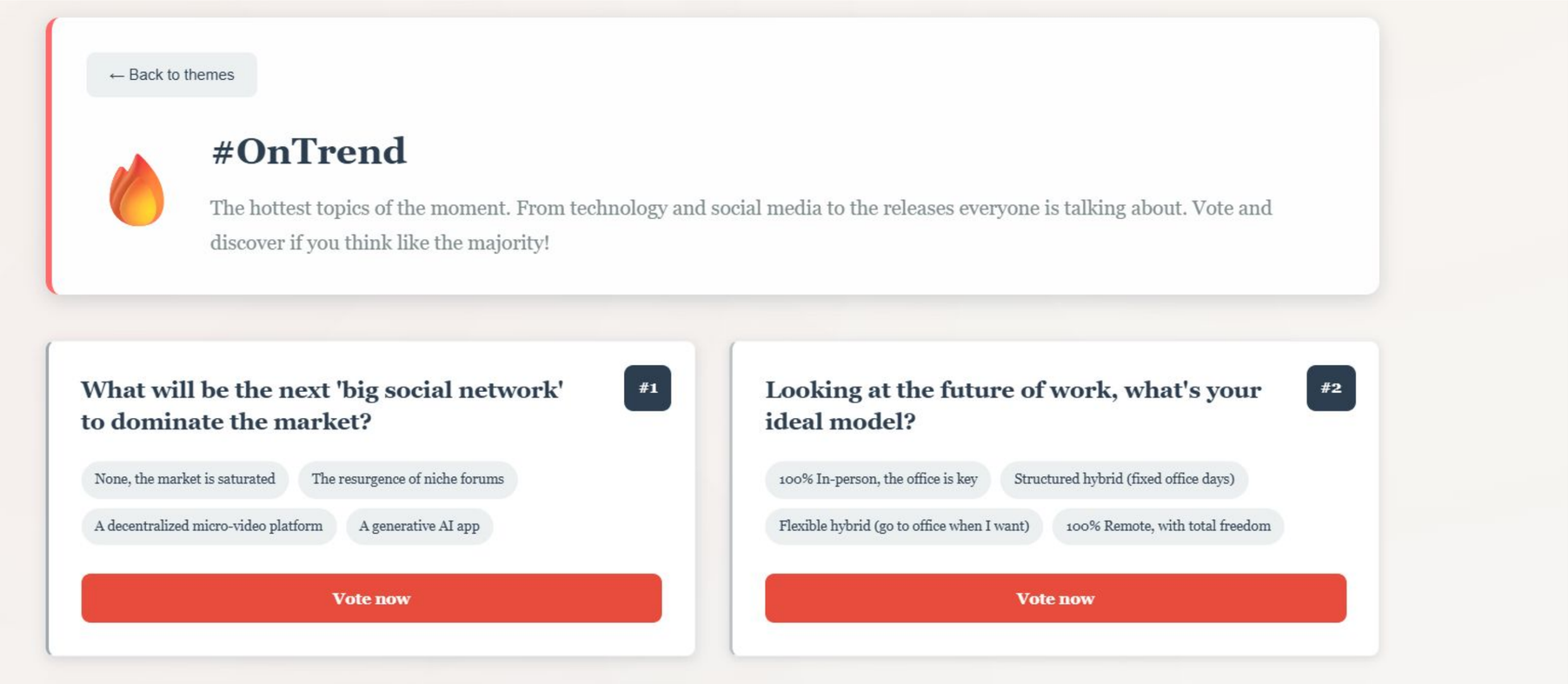
Technische Universität Berlin

Voting Live – Real-Time Updates



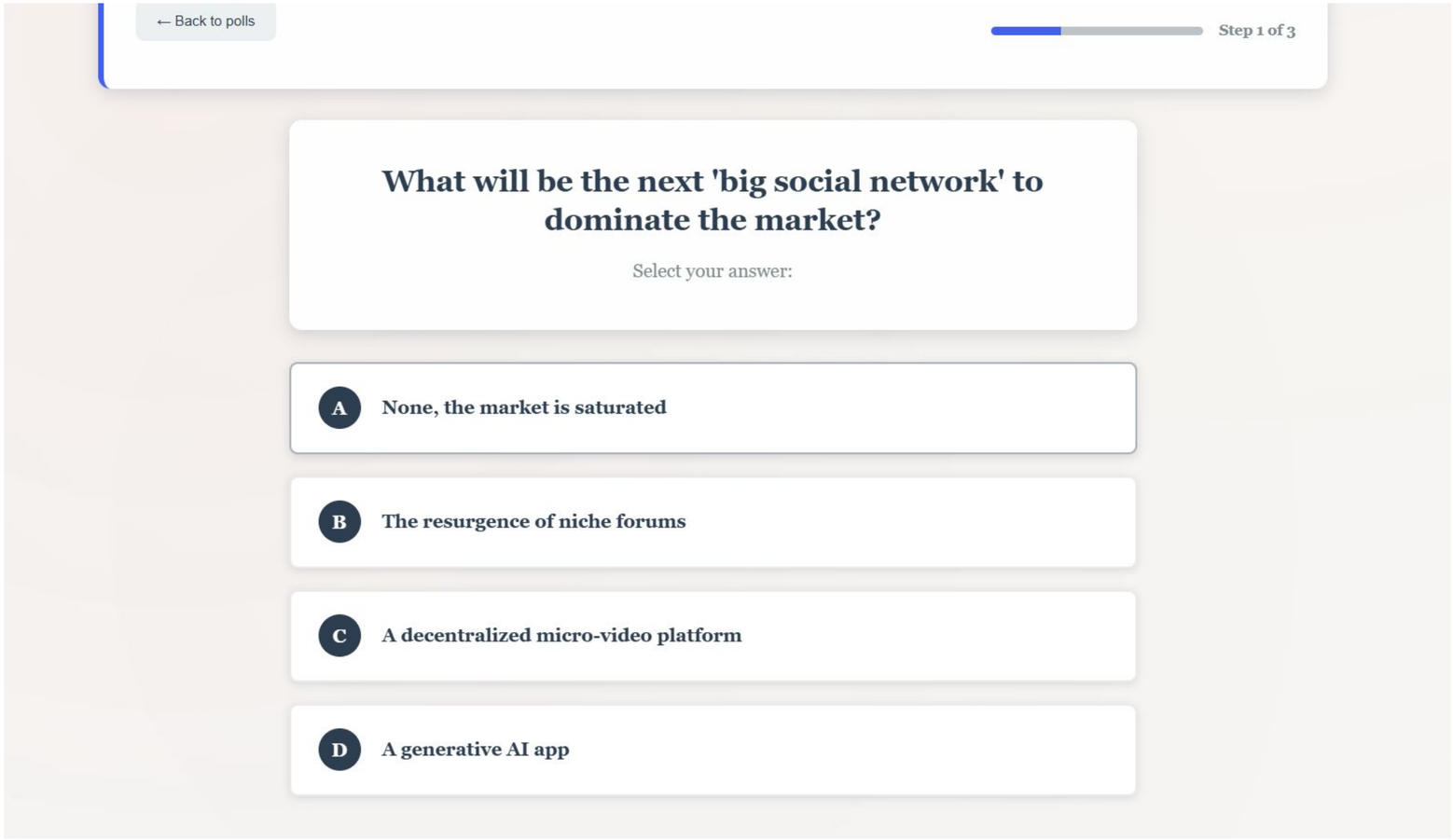
Technische Universität Berlin

Voting Live – Real-Time Updates



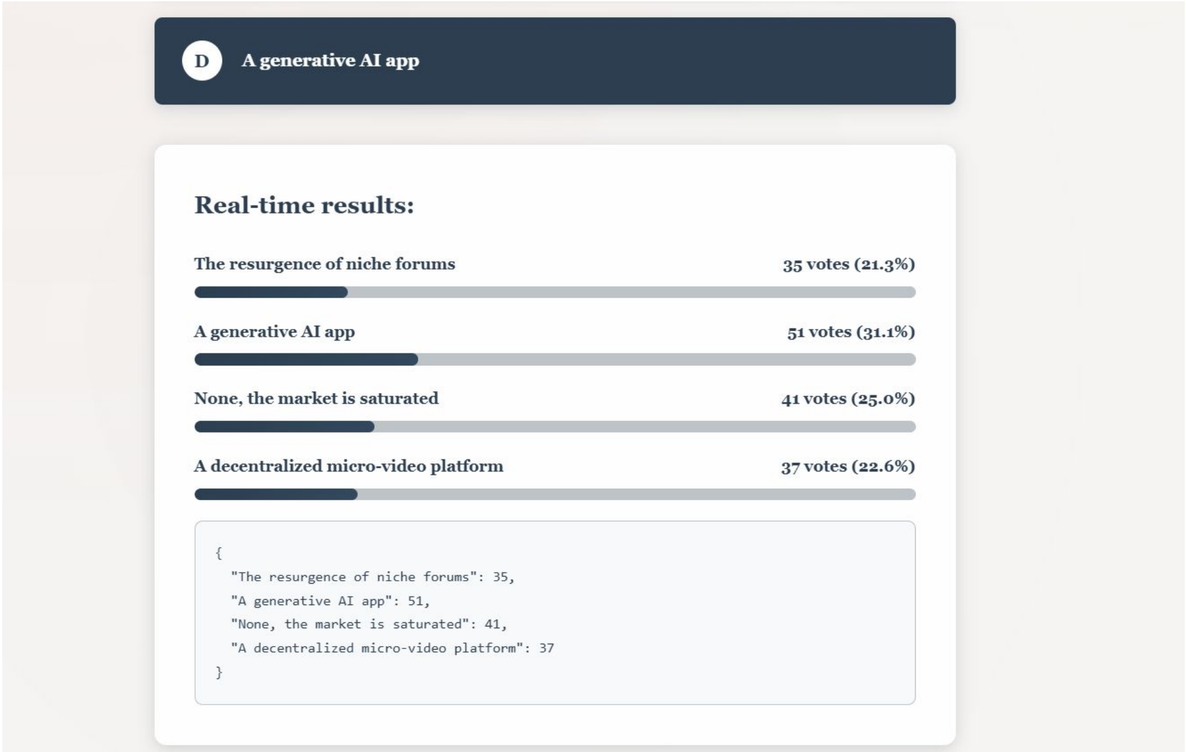
Technische Universität Berlin

Voting Live – Real-Time Updates



Technische Universität Berlin

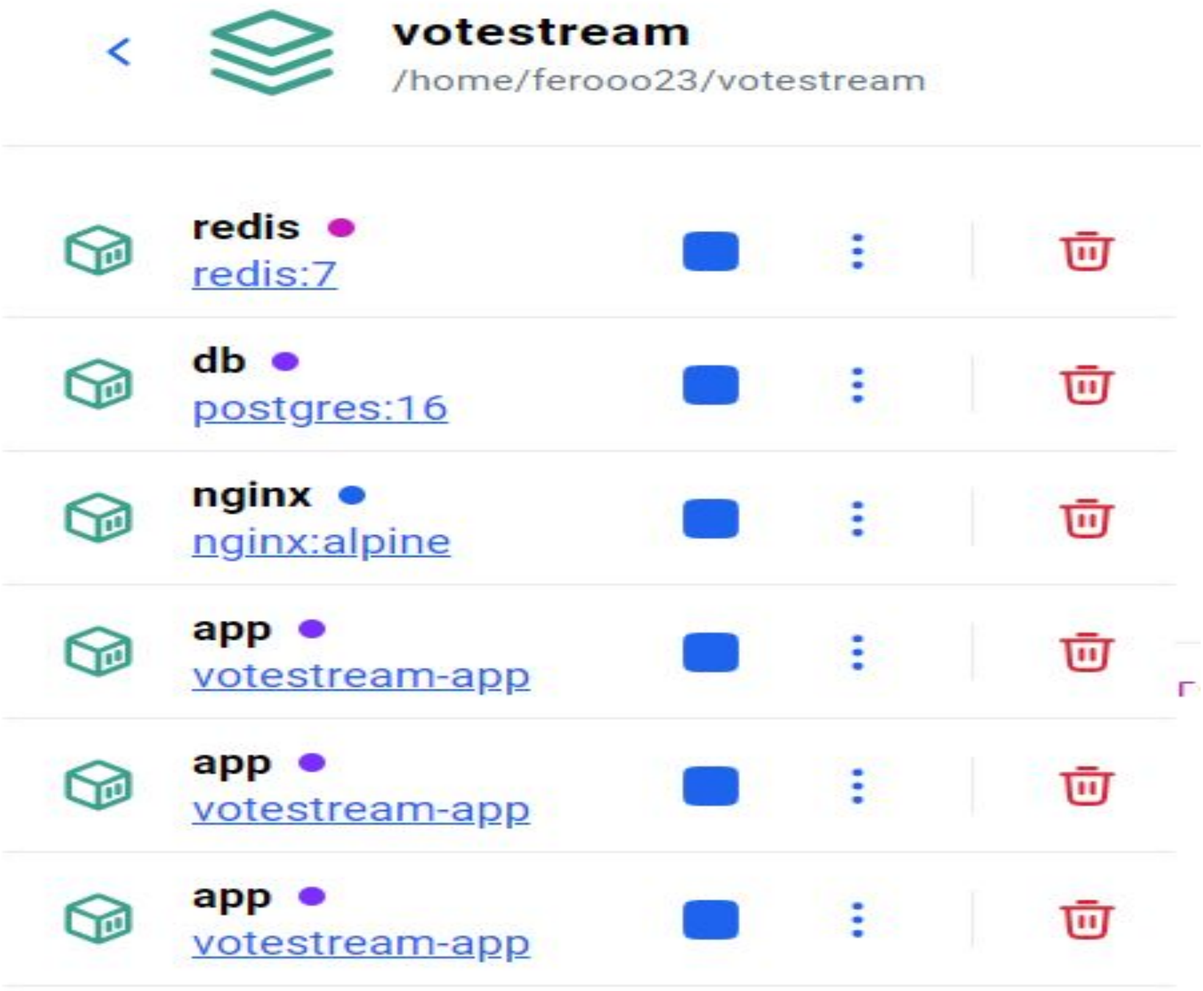
Voting Live – Real-Time Updates



- When a user selects an option, the vote is saved to PostgreSQL.
- Results are instantly updated for all active users via Redis Pub/Sub and WebSockets.
- This showcases real-time updates and concurrent connection handling.

Technische Universität Berlin

Scaling the System – Load Balancing



- Our stateless architecture allows multiple backend instances to run in parallel.
- Nginx acts as a load balancer and distributes incoming traffic evenly.
- Here we demonstrate multiple containers or Gunicorn workers handling requests concurrently.

Technische Universität Berlin

Scaling the System – Load Balancing


```
Phase 1: Testing rate limiting with rapid requests...
Response codes from 50 rapid requests:     50 429
```

```
⏳ Getting basic system status...
Health Status: {"detail":"Rate limit exceeded"}
Basic Stats: {"detail":"Rate limit exceeded"}
```

- A custom rate limiter restricts the number of requests per client to prevent abuse.
- A circuit breaker pattern protects Redis and the database from cascading failures under stress.
- This ensures system resilience and graceful degradation during high load.

Technische Universität Berlin

Conclusion

·VoteStream successfully meets all project requirements, demonstrating a robust, scalable, and resilient architecture.

·We have implemented state management, vertical/horizontal scaling, overload mitigation, caching, and asynchronous patterns.

·The project is fully containerized and easy to deploy.

🔗 GitHub Repository: https://github.com/ferooo23/votestream.git

Technische Universität Berlin