Introduction to Gabber

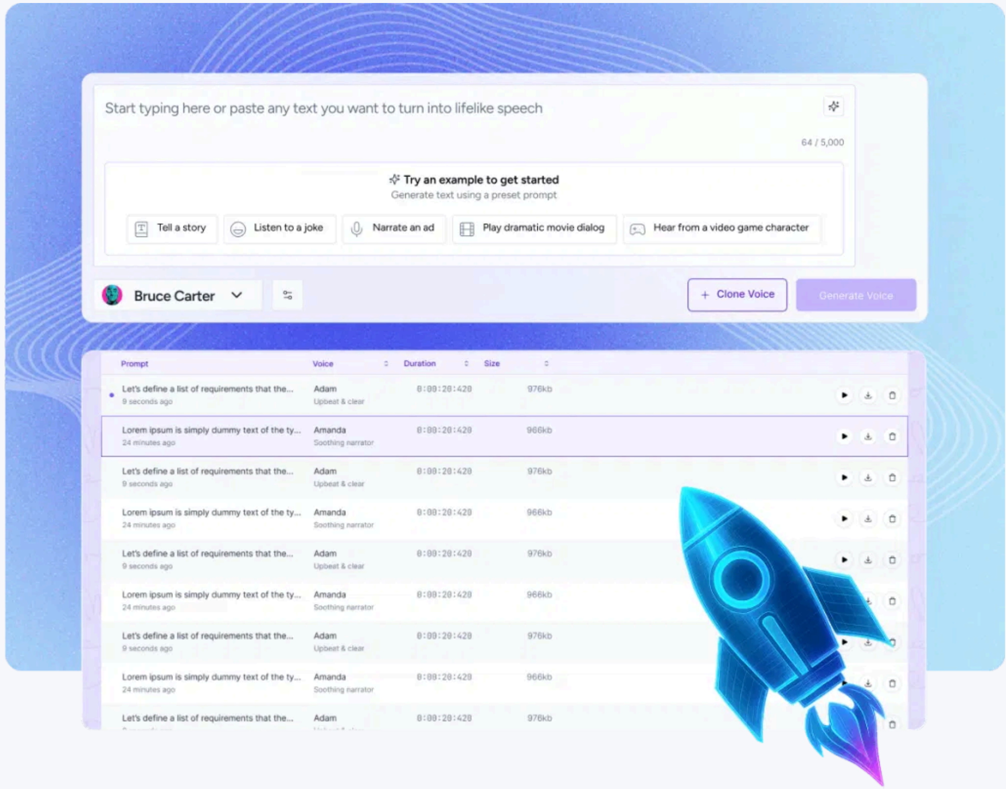
Gabber provides a **backend platform for AI personas** in applications such as:

- AI Companions
- Non-Player Characters (NPCs)
- AI Tutors

Key Focus Areas:

- Low latency for real-time interactions
- Continuity in voice conversations
- Cost-effective deployment (**$1 per hour** target)

Their approach leverages open-source tools and innovative techniques to make high-quality voice AI accessible to smaller teams.



The Challenge: Head-of-Line Silence

Initial Approach:

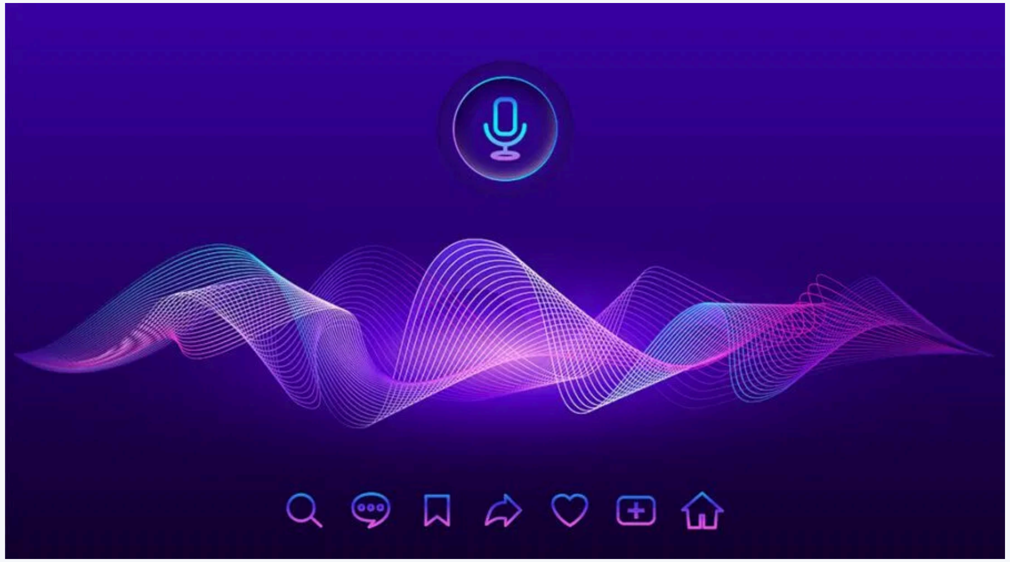
- Selected **Orpheus TTS** as the open-source text-to-speech model
- Chosen for its high-quality voice synthesis capabilities

The Problem:

- **600ms of "head-of-line silence"** at the beginning of each audio generation
- Significant negative impact on real-time conversation experience
- Unacceptable for applications requiring natural dialogue flow

Impact on User Experience:

Human conversation gap: ~200ms (natural)
Orpheus TTS initial gap: ~600ms (unnatural, noticeable delay)



Solution 1: LoRa Fine-tuning

The Approach:

- Applied **Low-Rank Adaptation (LoRa)** fine-tuning to the Orpheus TTS model
- Specifically targeted the elimination of the initial silence
- Required minimal training data compared to full model retraining

Key Benefits:

- Reduced initial latency from **600ms to ~100ms**
- Enabled creation of high-fidelity, emotive voice clones
- Required surprisingly small datasets for effective voice cloning

Impact:

The 100ms latency is within the range of natural human conversation gaps (~200ms), creating a much more natural dialogue experience.

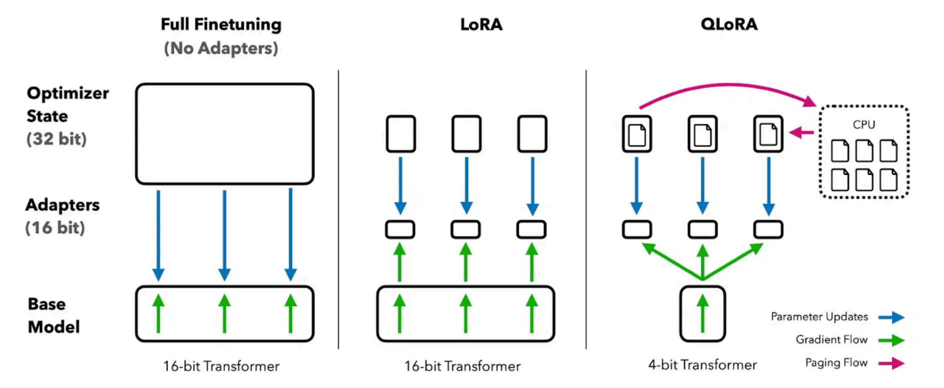


Figure 1: Different finetuning methods and their memory requirements. QLoRA improves over LoRA by quantizing the transformer model to 4-bit precision and using paged optimizers to handle memory spikes.

Solution 2: Efficient Inference

Hardware & Software Stack:

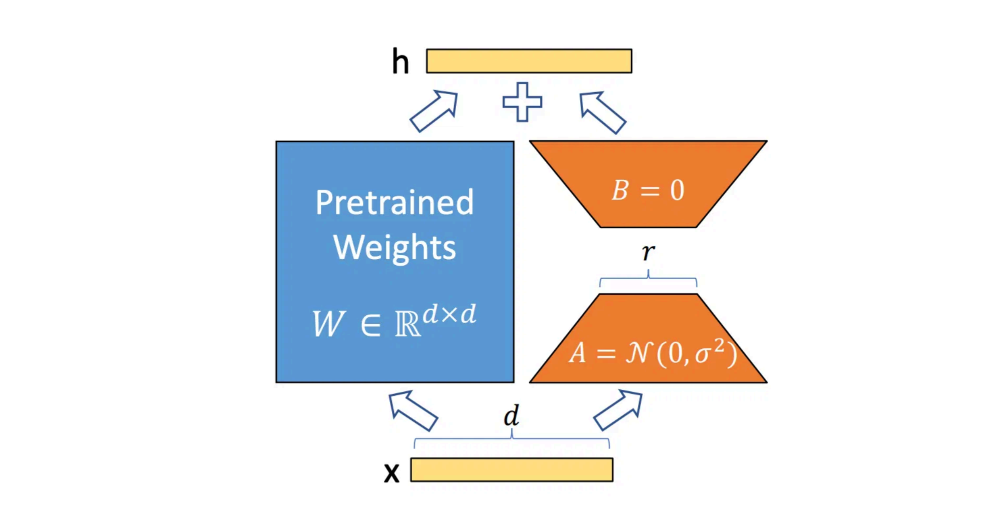
- Utilized **vLLM** inference engine on L40S GPUs
- Implemented **FP8 dynamic quantization** for optimized performance
- Enabled batch inference with multiple different LoRAs simultaneously

Performance Achievements:

- Achieved **over 100 tokens per second** processing speed
- Performance exceeds real-time requirements (faster than human speech)
- Maintained high quality while reducing computational requirements

Cost-Efficiency Impact:

The optimized inference stack is a key factor in achieving the $1/hour cost target while maintaining high-quality output.



Solution 3: Load Balancing

The Challenge:

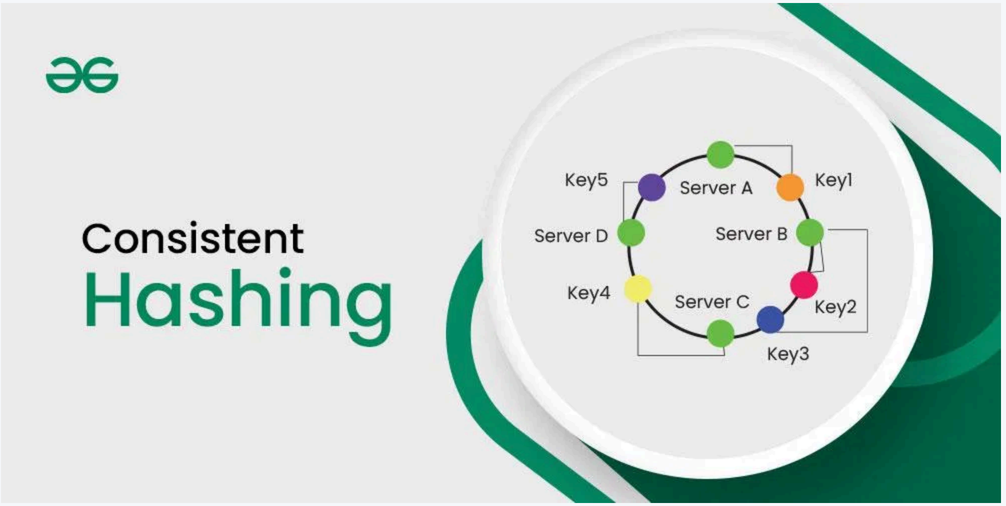
- Loading LoRAs on new servers causes significant latency spikes
- Traditional load balancing would distribute requests across all servers
- Need to maintain session continuity with specific voice models

The Solution:

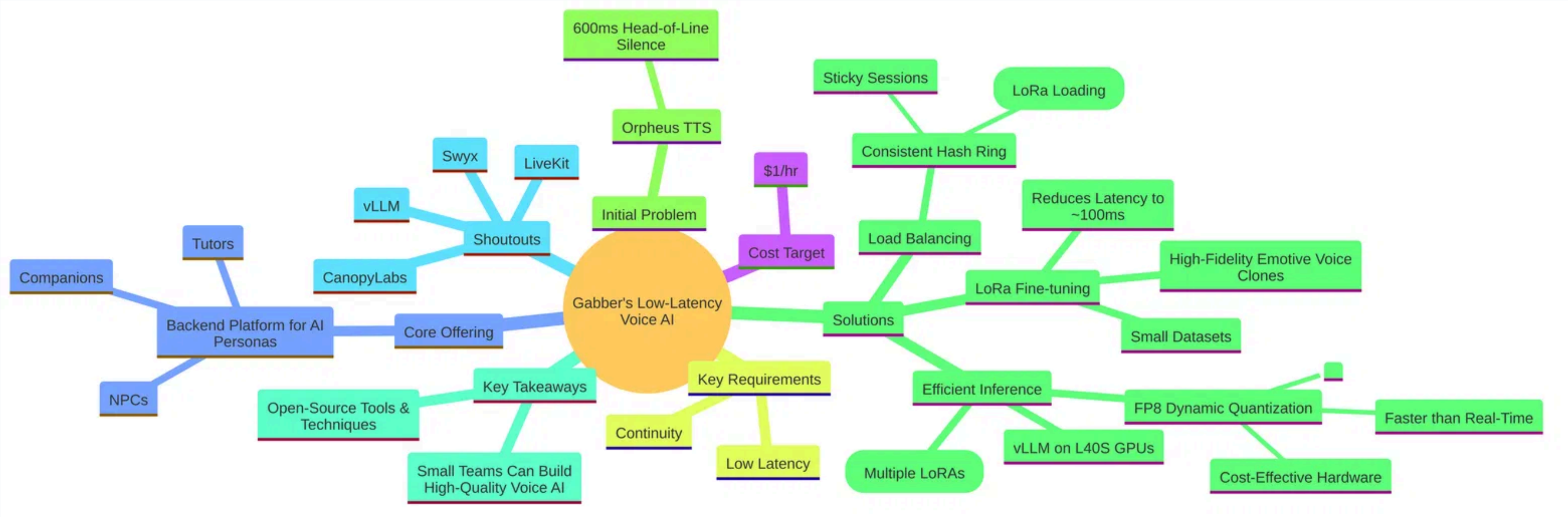
- Implemented a **consistent hash ring** for load distribution
- Created **sticky sessions** to maintain continuity
- Ensures sessions with specific cloned voices (LoRAs) remain on the same server

Key Benefit:

Once a session with a specific voice clone starts, it stays on the server where that LoRA is already loaded, eliminating mid-conversation latency spikes.



Gabber's Architecture Overview



The complete architecture integrates all components to deliver **low-latency, high-quality voice AI at $1/hour**.

Voice Generation

LoRA-fine-tuned Orpheus TTS model with reduced initial silence (100ms) and high-fidelity voice cloning capabilities.

Inference Optimization

vLLM on L40S GPUs with FP8 dynamic quantization achieving >100 tokens/second, faster than real-time speech.

Load Distribution

Consistent hash ring creating sticky sessions to keep voice models (LoRAs) on the same server throughout a conversation.

Key Takeaways

Lessons from Gabber's Implementation:

- **Open-source tools** can deliver professional-grade voice AI
- **LoRA fine-tuning** enables efficient customization with minimal data
- **Latency optimization** is critical for natural-feeling AI conversations
- **Thoughtful architecture** can dramatically reduce operational costs

The Big Picture:

Small teams can now build and host high-quality, low-latency voice AI at affordable costs ($1/hour), making advanced voice technology accessible to a wider range of applications and developers.

Swyx CanopyLabs LiveKit vLLM