Revolutionary Voice Enhancement in Real-Time Communications with GPU

Davit Baghdasaryan, CEO, 2Hz
Arto Minasyan, CTO, 2Hz

ARTIFICIAL INTELLIGENCE

Real-Time Noise Suppression Using Deep Learning

By Davit Baghdasaryan | October 31, 2018 Tags: 2hz.ai, Cloud Services, CUDA, Deep Learning, edge computing, machine learning and AI, noise suppression, Telecommunications, telecoms

Mute Background Noises

Voice Quality with Deep Learning

- Mute Background Noise
- Mute Everyone Except Me
- Remove Room Echo
- High Resolution Voice Everywhere

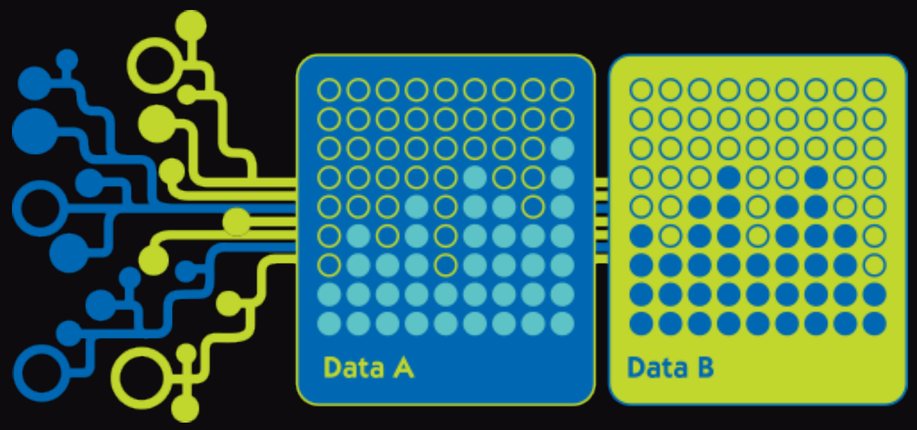
Real-Time Noise Suppression with Deep Learning

Traditional Noise Cancellation

- Requires 2-4 mics
- Runs on edge device
- Cancels only limited noises
- Outbound only

Deep Learning powered Noise Cancellation

Train krispNet
Deep Neural Network



Background Noises

Clean Human Speeches

- No dependency on mics
- Bi-directional
- Cancels all noise types
- Runs everywhere - on device and in the cloud

How to Measure Voice Quality?

Industry Standards

- Academia - PESQ, Subjective
- Industry - 3QUEST (Speech MOS, Noise MOS, Global MOS)
- Skype Audio Test and 3GPP TS 26.131 specifications

Audio Lab

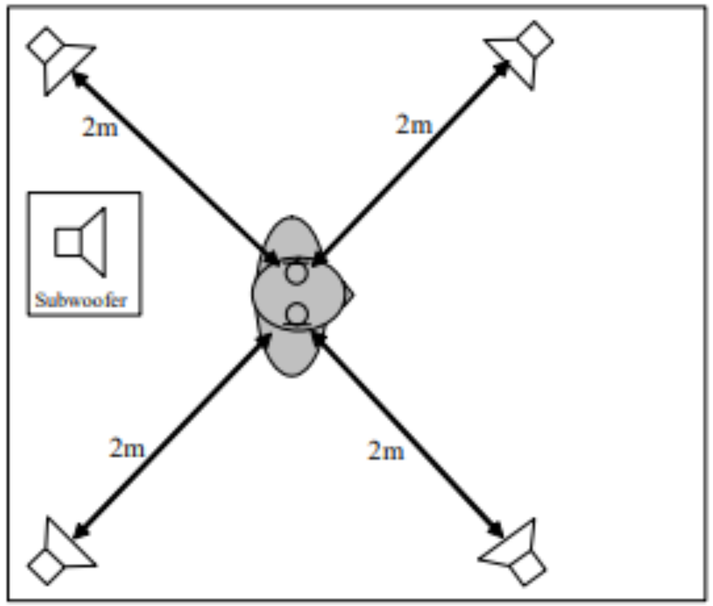
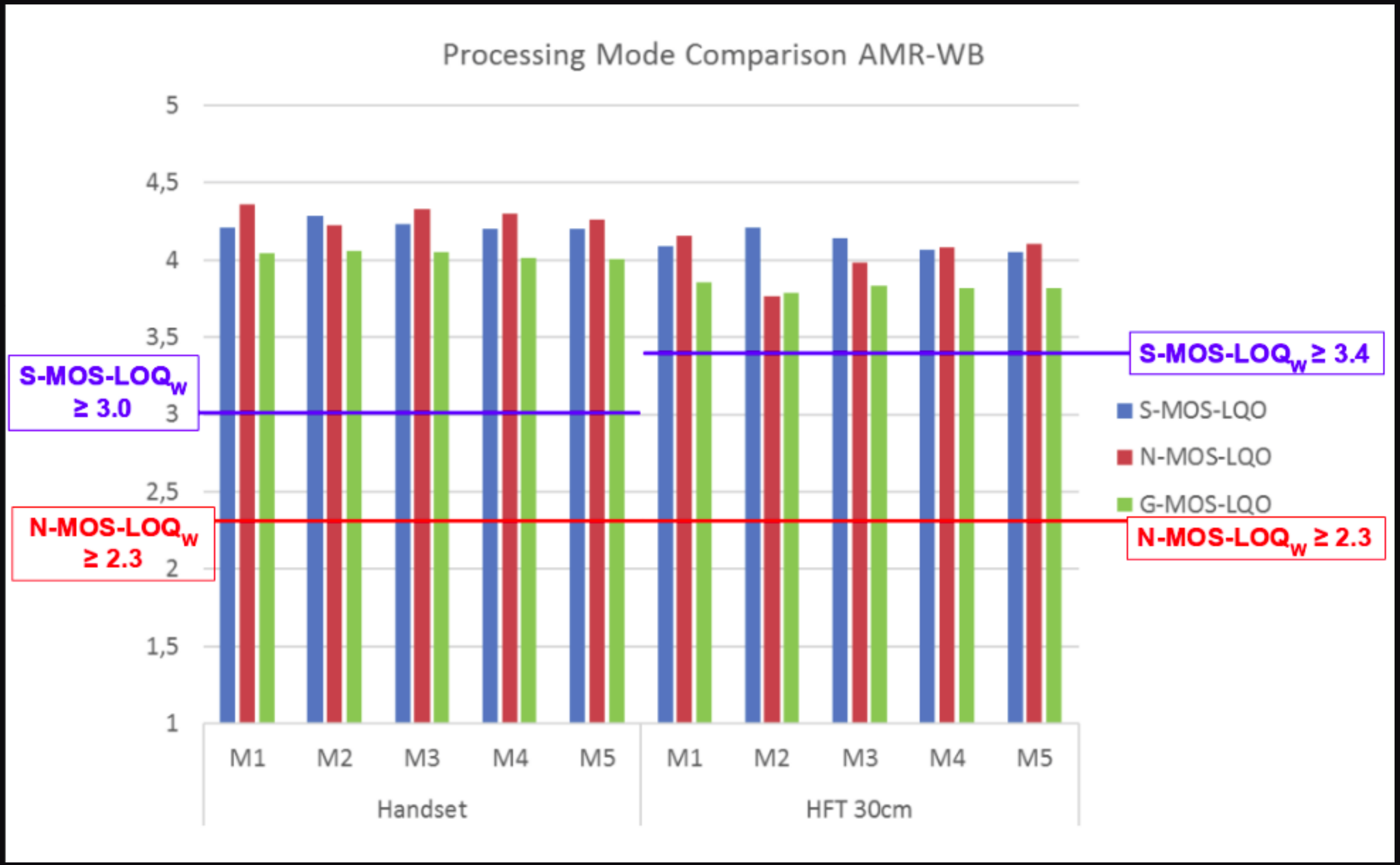


Figure 2: Loudspeaker arrangement in standard office rooms


Processing Mode Comparison AMR-WB
5
4,5
4
3,5
3
2,5
2
1,5
1
S-MOS-LOQ$_W$ ≥ 3.0
N-MOS-LOQ$_W$ ≥ 2.3
S-MOS-LOQ$_W$ ≥ 3.4
S-MOS-LQO
N-MOS-LQO
G-MOS-LQO
M1
M2
M3
M4
M5
M1
M2
M3
M4
M5
Handset
HFT 30cm

krisp.ai

App for Mac

ONE BUTTON, NO NOISE.

Mute the background noise during your calls

Seamlessly Integrates in Conferencing Apps

Supports any Microphone or Headset

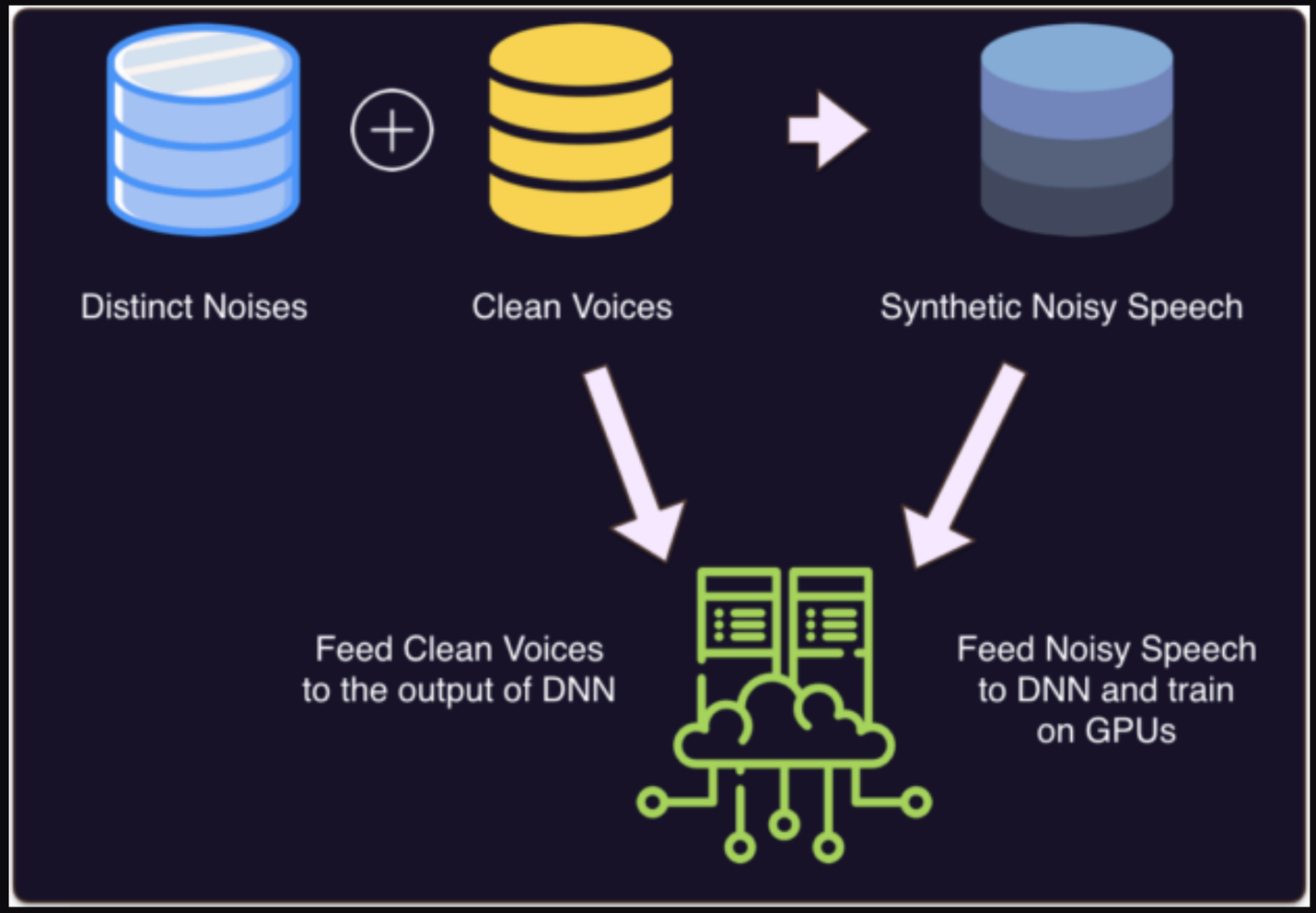

NORTH
COMMUNITY
BANK

krisp.ai

Best Product in Audio/Voice 2018

Training and Inference

Training Process



Training Data

- 2K distinct speakers - gender and age diverse distribution
- >10K distinct noises - babble, construction, traffic, cafeteria, office, etc
- 2000+ hours

Training on GPUs

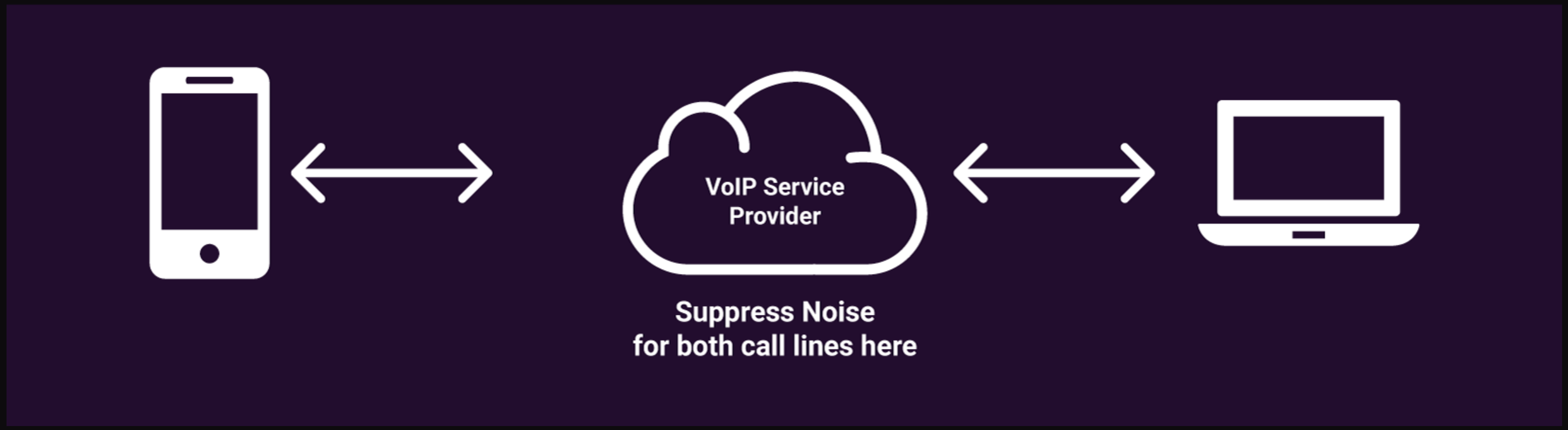
- All in Python
- Distributed TensorFlow
- Multiple in-house NVIDIA 1080ti. Takes a full week.
- p2.16xlarge in AWS. 16x NVIDIA K80

Inference

- Supports NVIDIA, Intel and ARM platforms
- All in C/C++. Sometimes ASM
- Smaller network (5x boost with some quality penalty)
- TensorRT boosts ~2x

Moving to the Cloud

Server-side Noise Cancellation



Latency Constraints

200ms end to end latency

Codecs and other DSP (10-80ms)

Network (varies)

DNN Compute (< 5ms)

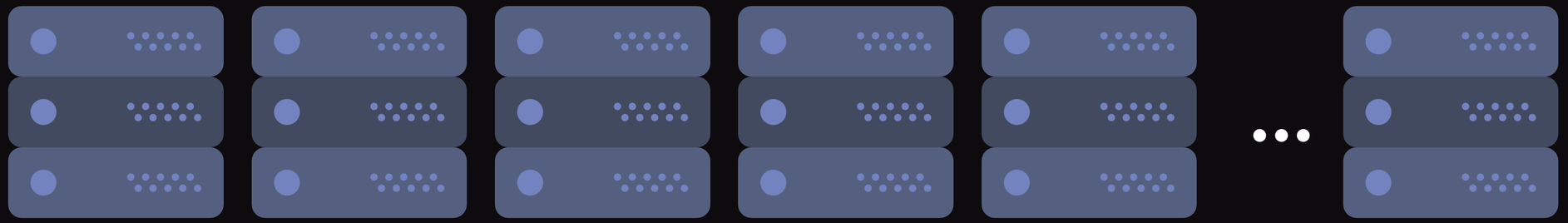
DNN Algorithmic (15ms)

< 20ms

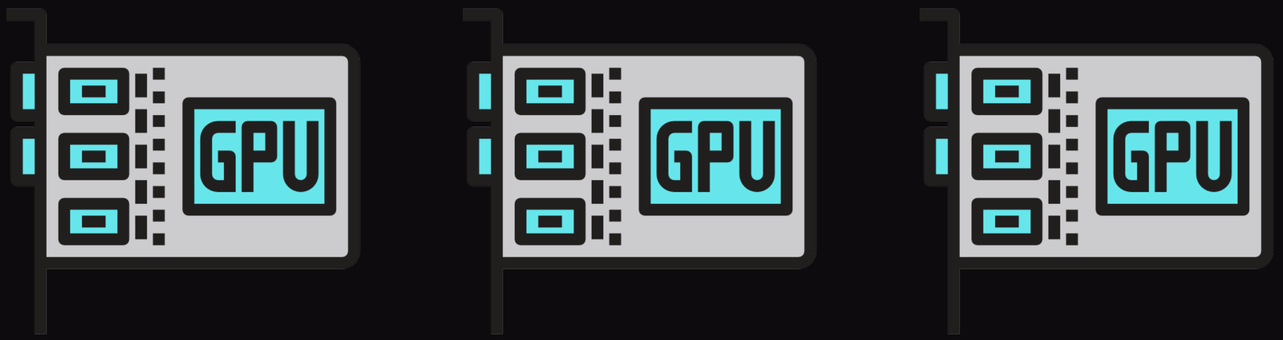
How do you scale to 100K+ concurrent streams with such latency constraints?

Ex. Discord processes 2.5M concurrent audio streams

CPU Servers

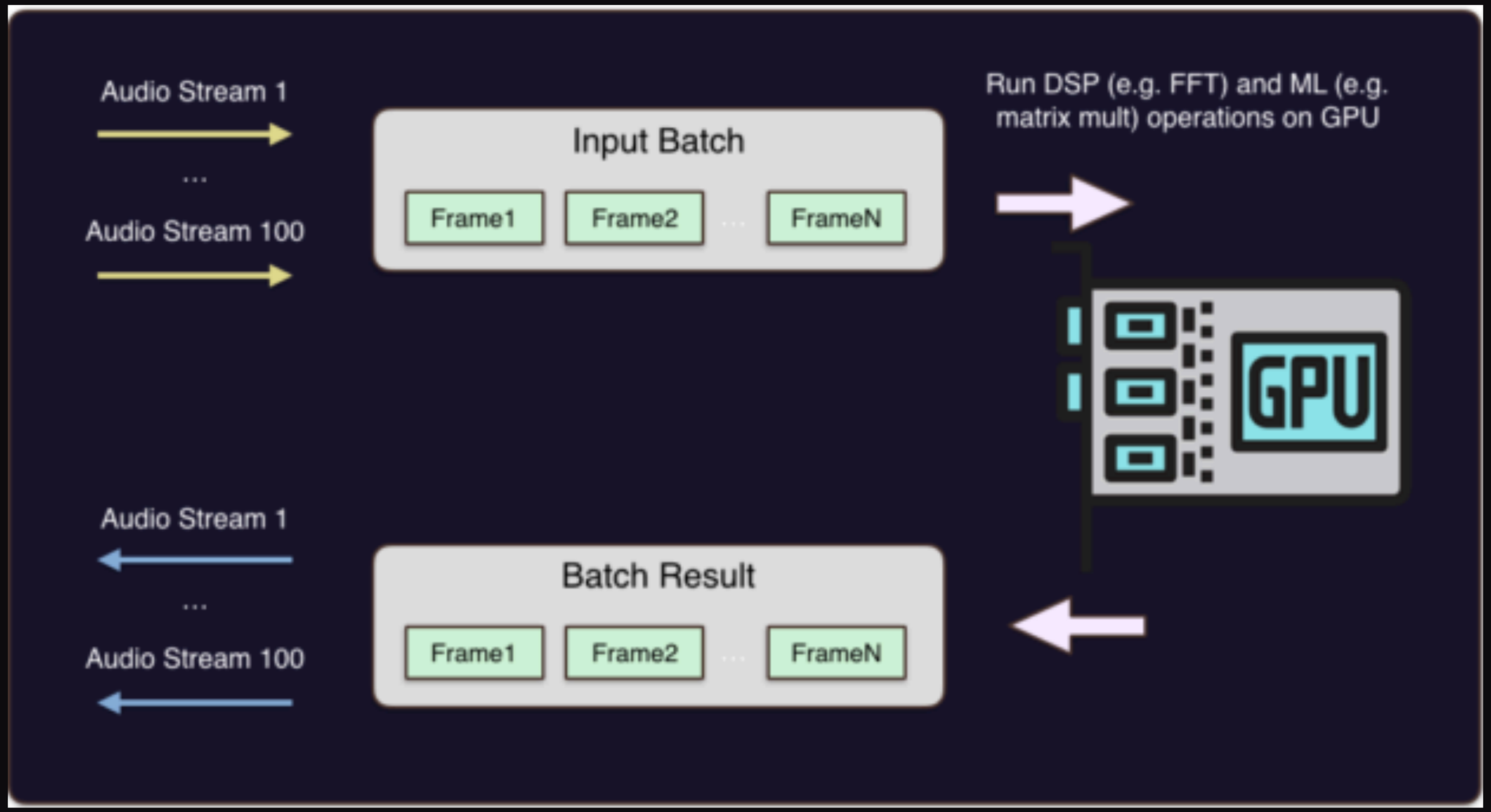


GPU Servers

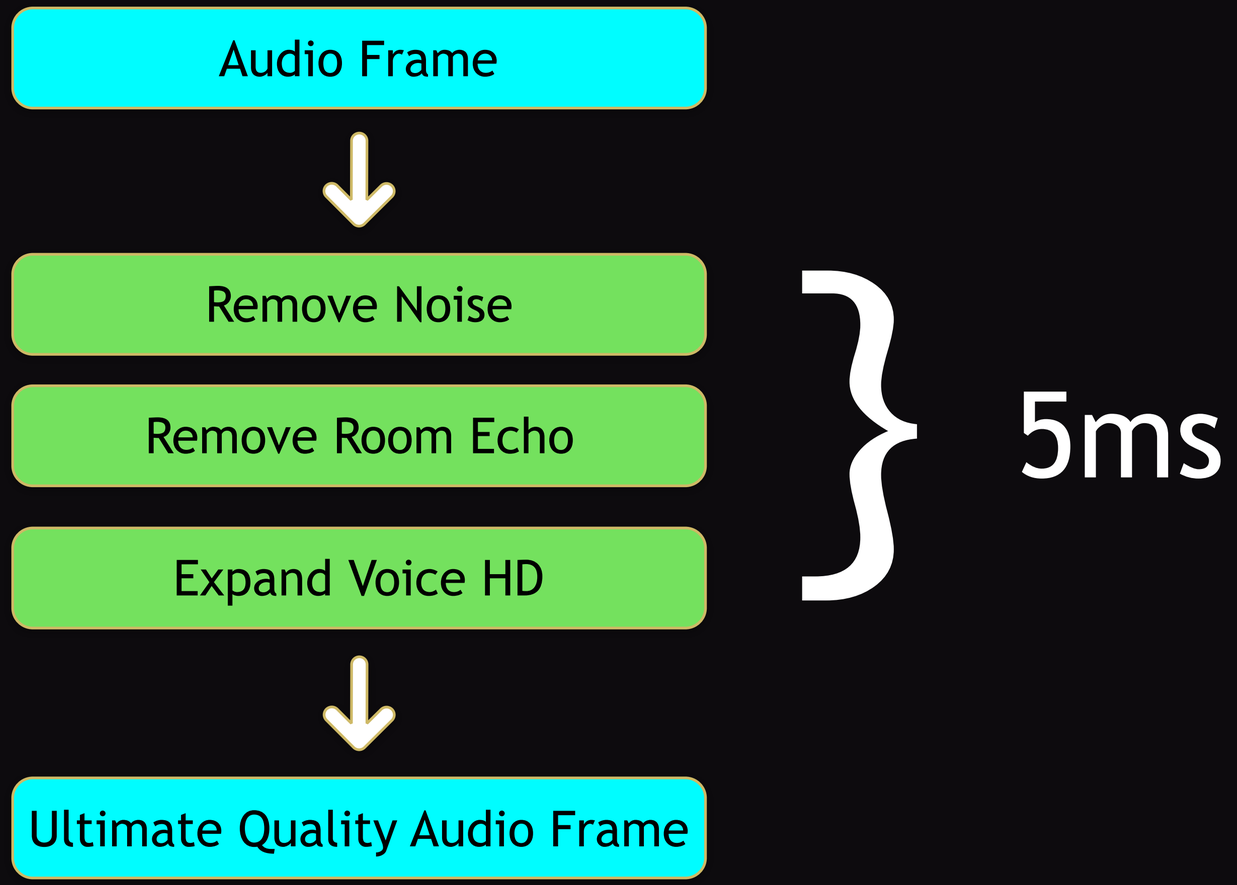


10x-20x less costly

Scalability with Batching

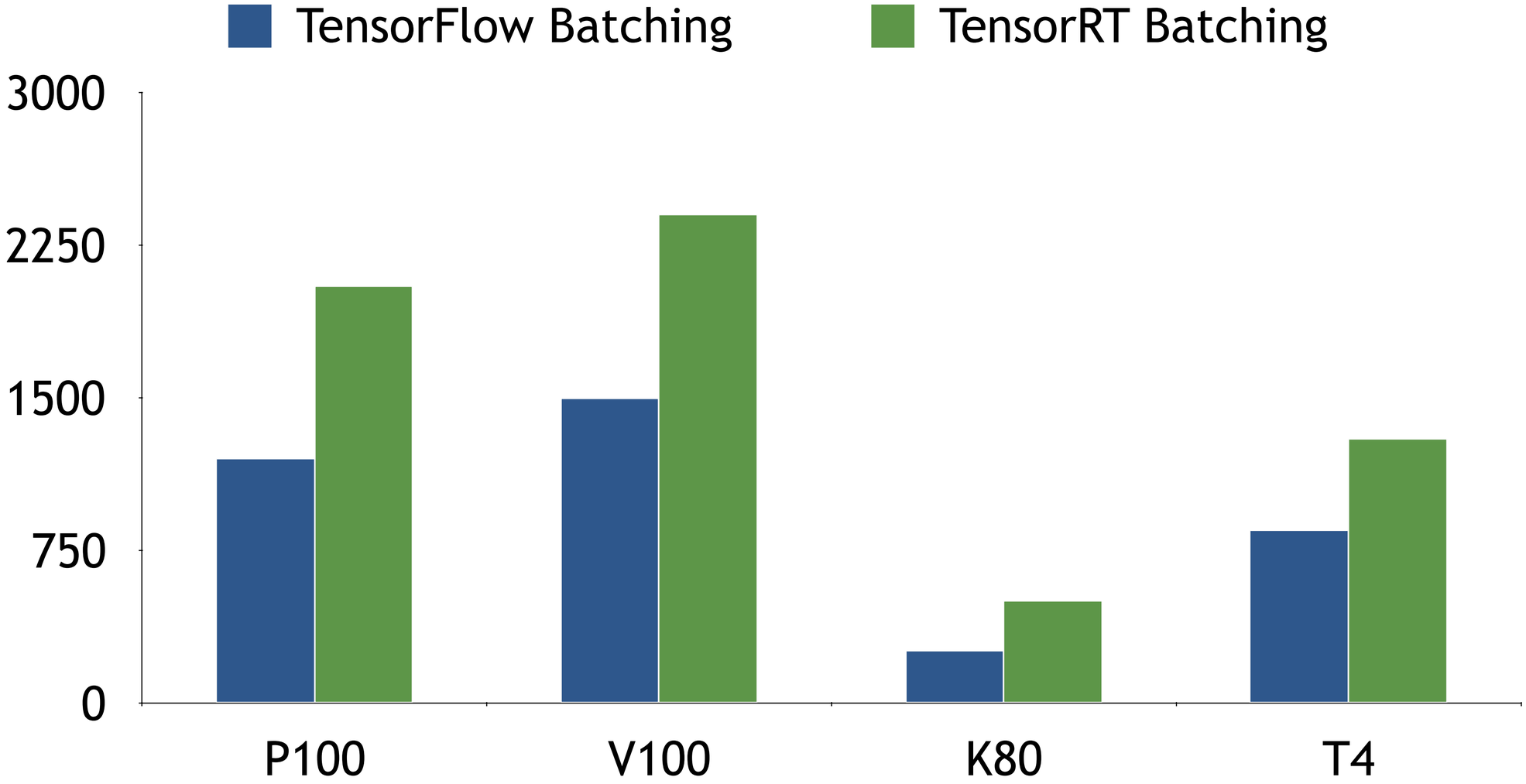


Ultimate Quality

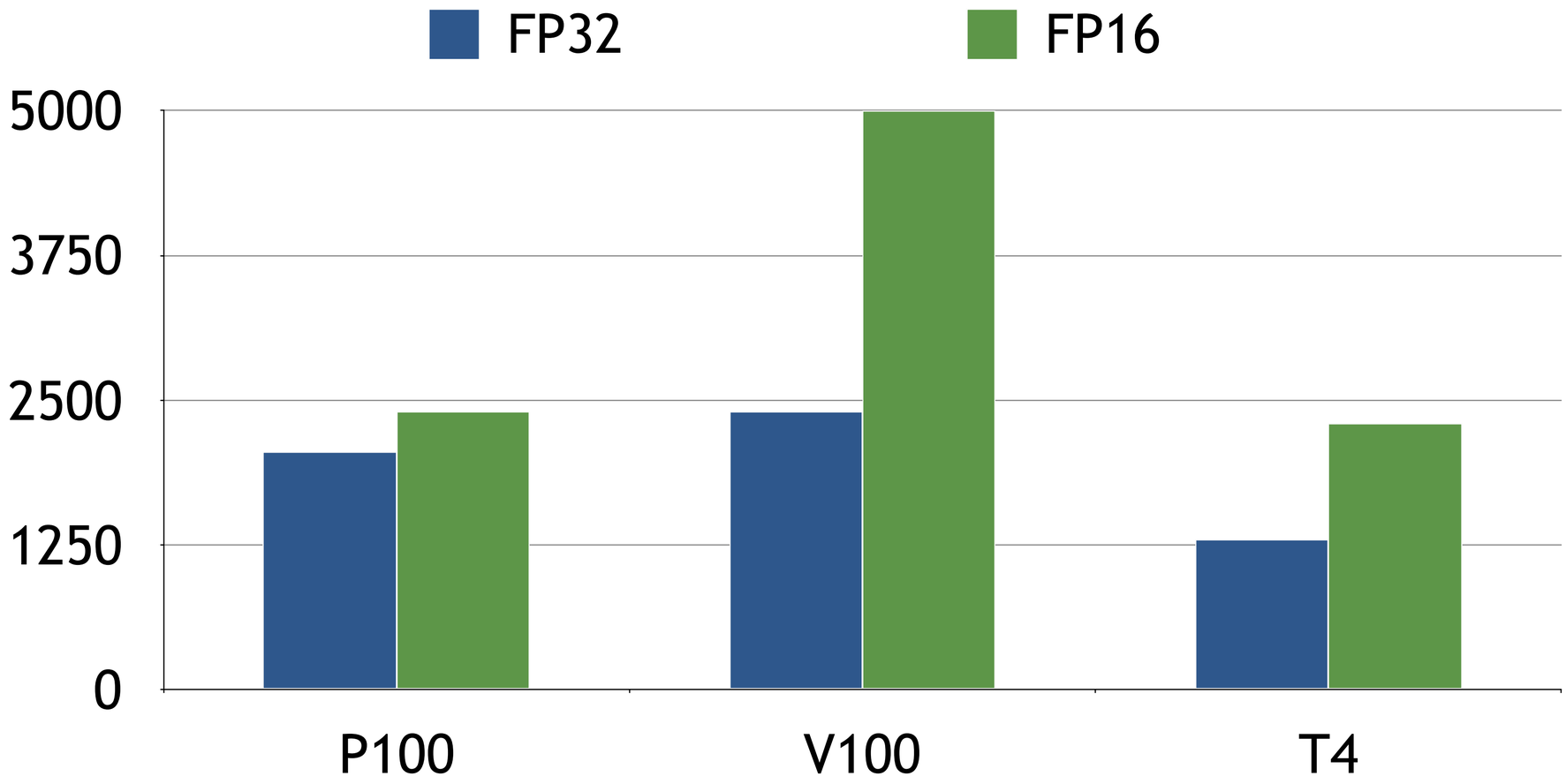


Maximum Quality and Scale with NVIDIA Tensor Cores

TensorRT is pretty awesome



T4 and V100 are both awesome



Key Takeaways

1. Voice Quality Enhancement is moving to the Cloud

2. For large scale deployments we need GPUs

3. T4 and V100 GPUs are most efficient for this

Thank You!

Booth #247