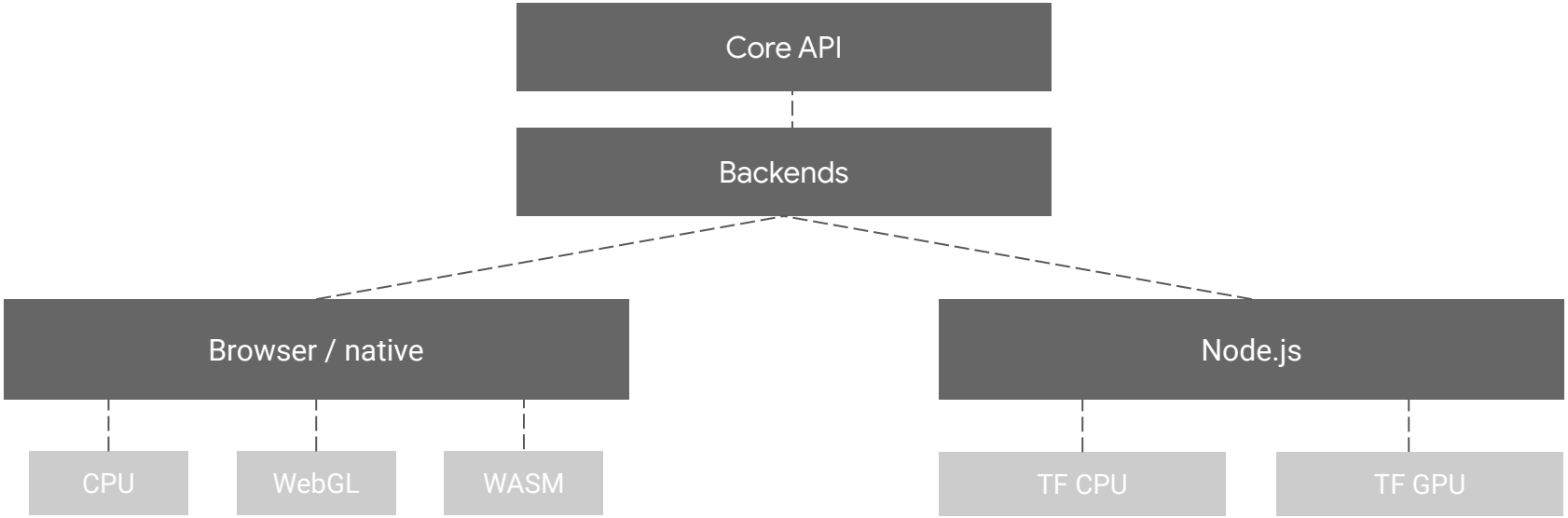
Fast client-side ML with TensorFlow.js

Ann Yuan

Software Engineer - **TensorFlow.js**, Google

What is TensorFlow.js?



Performance overview

	WebGL	WASM	WASM+SIMD	Plain JS
iPhone XS	18.1	140		426.4
Pixel 3	77.3	266.2		2345.2
Desktop Linux	17.1	91.5	61.9	1049
Desktop Windows	41.6	123.1	37.2	1117
MacBook Pro 2018	19.6	98.4	30.2	893.5

Inference times for **MobileNet** in ms.

Performance overview

	WebGL	WASM	WASM+SIMD	Plain JS
iPhone XS	10.5	21.4		176.9
Pixel 3	31.8	40.7		535.2
Desktop Linux	12.7	12.6		249.5
Desktop Windows	7.1	16.2	7.5	270.9
MacBook Pro 2018	22.7	13.6	7.9	209.1

Inference times for **FaceDetector** in ms.

The WebAssembly backend

How it works:

- Language: **C++**
- Compiler: **Emscripten**
- Accelerator: **XNNPACK**

WebAssembly: our wishlist

- Broader SIMD, SharedArrayBuffer support
- Wider SIMD
- Quasi fused multiply-add
- Pseudo minimum / maximum
- ES6 module support

The WebGL backend

How it works:

- Data: **GPU textures**
- Computation: **Fragment shaders**

WebGL: our wishlist

- Improved portability
- Tools for memory management
- Callbacks for data download

The WebGPU backend

How it works:

- Data: **Storage buffers**
- Computation: **Compute shaders**

WebGPU: Performance

	WebGPU	WebGL
Discrete GPU (Radeon Pro 555X)	41.7	51.1
Integrated GPU (Intel UHD Graphics 630)	119.5	106.5

Inference times for **PoseNet** (ResNet50 architecture) in ms.

Future web standards wishlist

- Portability
- Tools for memory management
- Tools for model obfuscation
- Detailed profiling
- Low-level API