Web前沿技术论坛

WebGPU和Web AI

顾扬

英特尔“Web图形和Web AI” 团队经理

2023年6月19日

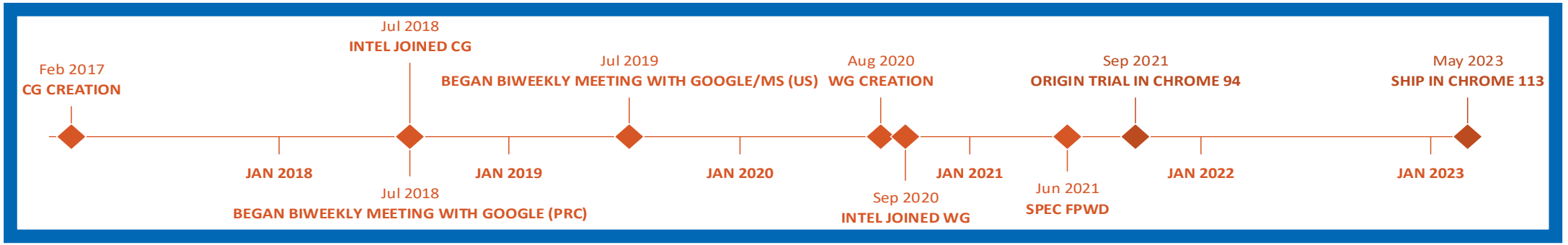
Agenda

- WebGPU
- Web AI
- 英特尔“Web图形和Web AI”团队


WA

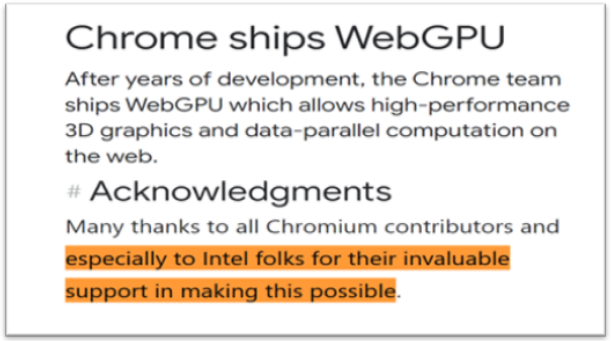
WebGPU

WebGPU



- 标准
  - 最复杂Web标准，包含100+ APIs和一个新的着色器语言WGSL
  - 贡献来自主要浏览器厂商 (Google，Apple，Mozilla和Microsoft)，其它公司(特别是英特尔)和个人
  - API和WGSL都接近正式发布
- 浏览器实现
  - 2023年5月2日，Chrome在M113正式发布 WebGPU!
  - Firefox很可能今年发布，Safari也在积极开发
- 先行者
  - AI框架: TensorFlow.js, ONNX Runtime, TVM，IREE
  - AI应用: Google Meet, Adobe Photoshop Web, Zoom
  - 游戏引擎: Unity, Unreal, Cocos, PlayCanvas, Construct3
  - 渲染框架: Three.js, Babylon.js, Orillusion
  - 其它: Snap, Node.js, Deno, Google Earth, Sketchfab

终于发布啦！



Chrome ships WebGPU

After years of development, the Chrome team ships WebGPU which allows high-performance 3D graphics and data-parallel computation on the web.

Acknowledgments

Many thanks to all Chromium contributors and especially to Intel folks for their invaluable support in making this possible.

Sundar Pichai @sundarpichai · Apr 7

WebGPU in Chromium 113, excited for the web and great to see this ship to stable!

Corentin Wallez @DaKangz · Apr 6

This feels unreal! After more than 6 years working on WebGPU, it's getting released in Chromium 113, in stable and without flags! It only took a bit longer than the 2 year adventure we initially thought it would be Read more about it here developer.chrome.com/blog/webgpu-re...

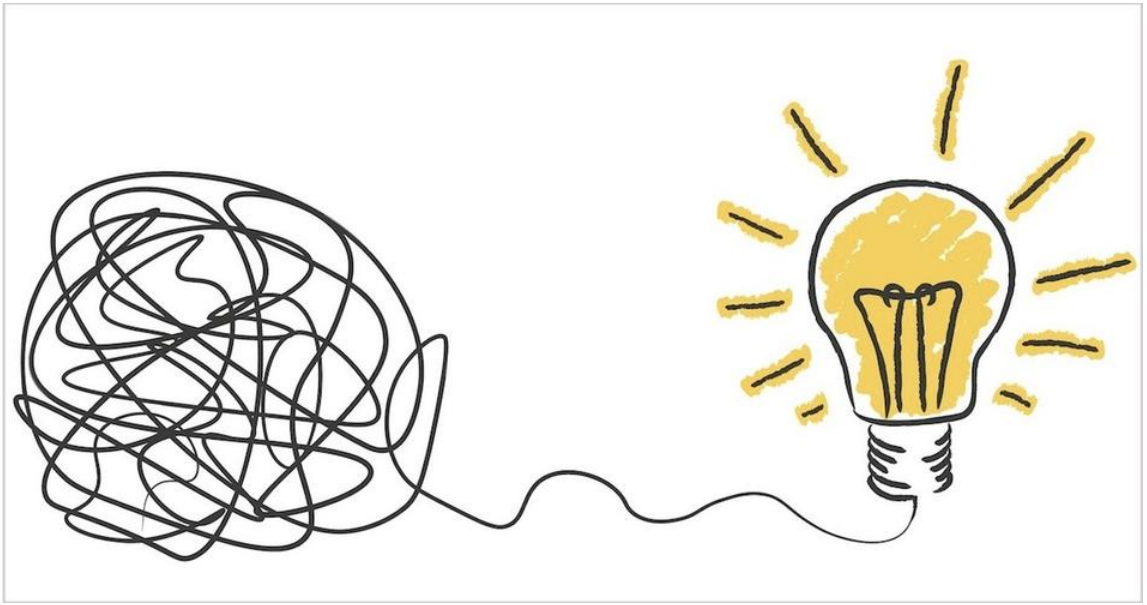
Show this thread

67 183 1,440 467.1K

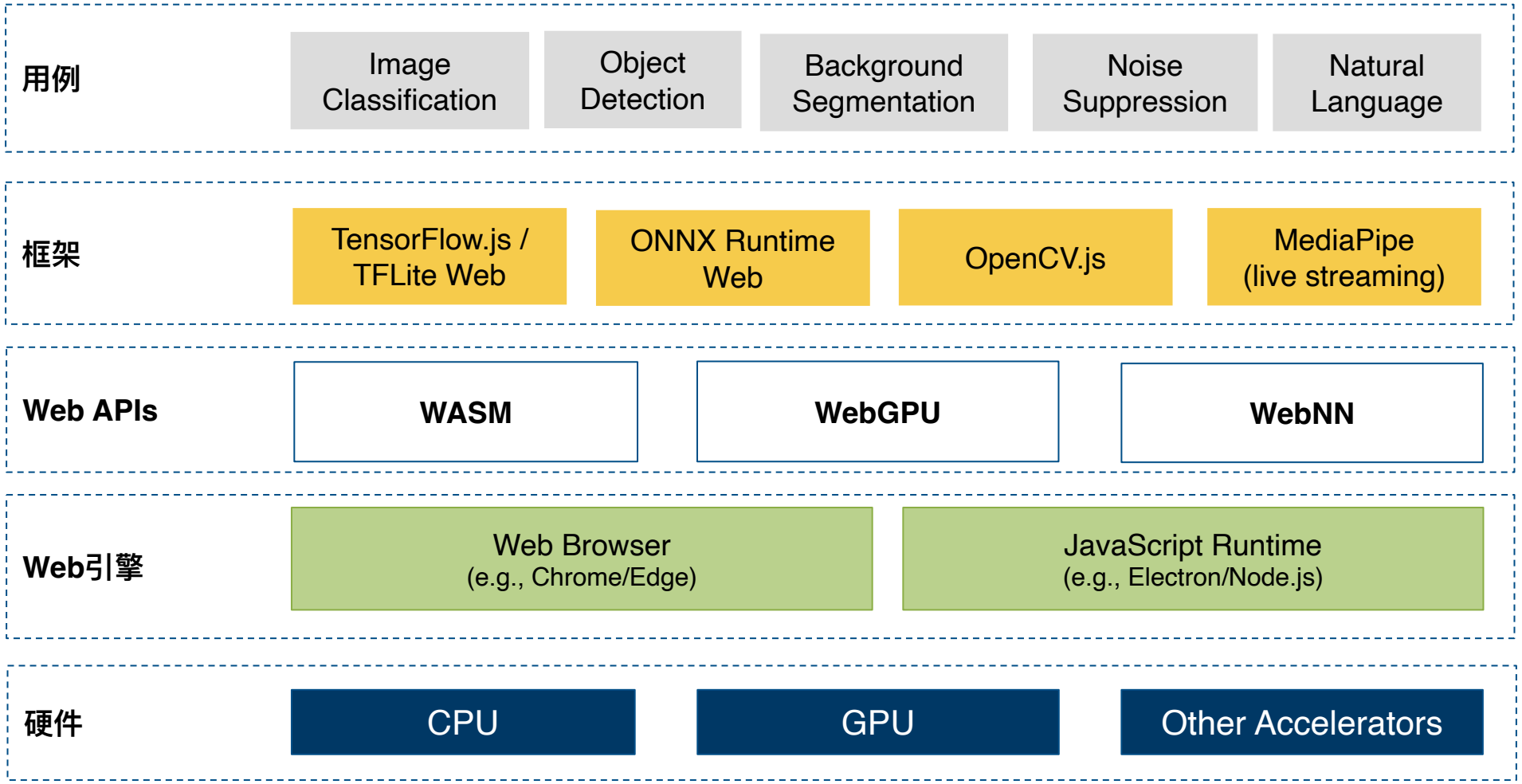
Web AI

Web AI 的一些总体想法

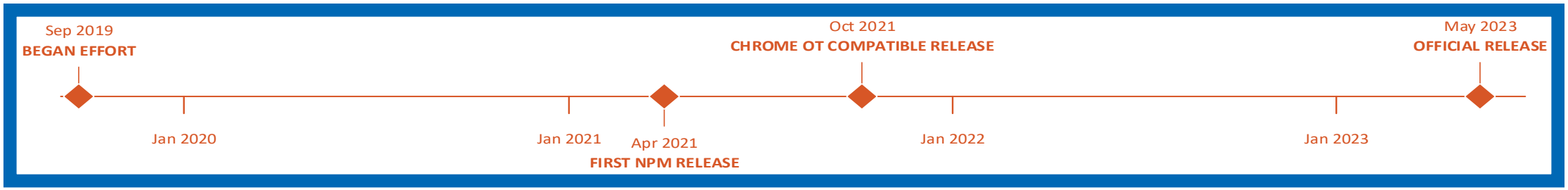
- 考虑隐私，成本，响应时间等因素，Client AI 已经是大势所趋
- Web AI能帮助Client AI更近一步，让每个人都能享用
- Web AI并不等同于轻量级模型。大模型，特别是LLM，有强烈的需求
- AI本身还在高速发展中，对于Web来说，成熟的方案还为时尚早
- Web AI需要接入Native的生态，一同发展



Web AI 架构图

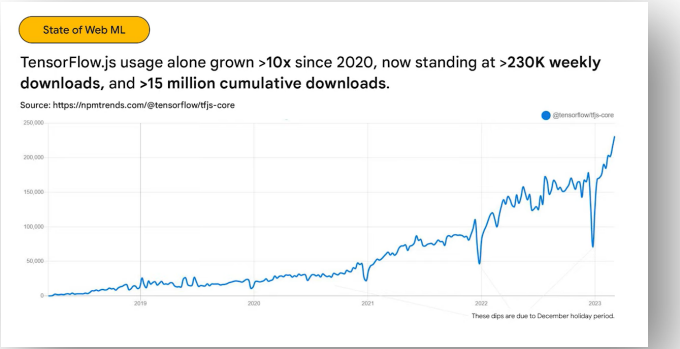

用例
Image Classification
Object Detection
Background Segmentation
Noise Suppression
Natural Language
框架
TensorFlow.js / TFLite Web
ONNX Runtime Web
OpenCV.js
MediaPipe (live streaming)
Web APIs
WASM
WebGPU
WebNN
Web引擎
Web Browser (e.g., Chrome/Edge)
JavaScript Runtime (e.g., Electron/Node.js)
硬件
CPU
GPU
Other Accelerators

TensorFlow.js WebGPU



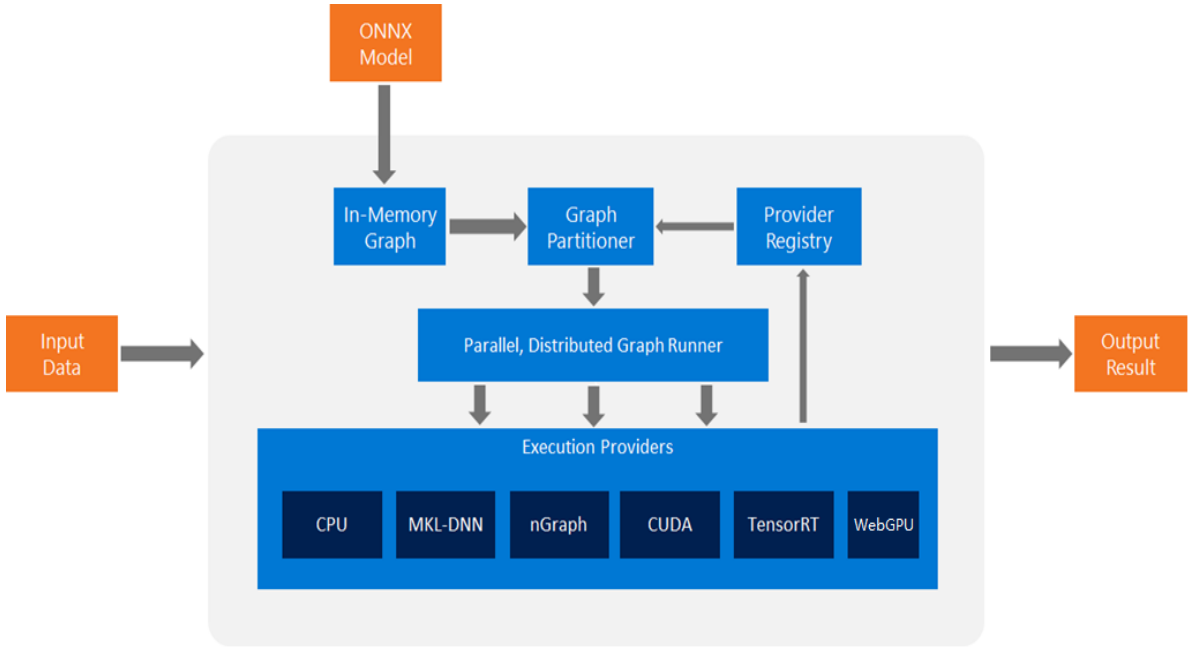
- Web上最流行的AI框架之一，每周230K下载
- 2023年5月18日发布第一个正式WebGPU版本 4.6.0
- 支持160/174 kernels
- 性能全面超越WebGL (一开始只有1/3)。除了很小的模型，比WASM性能好很多
- 500+ PR

我们也发布啦！



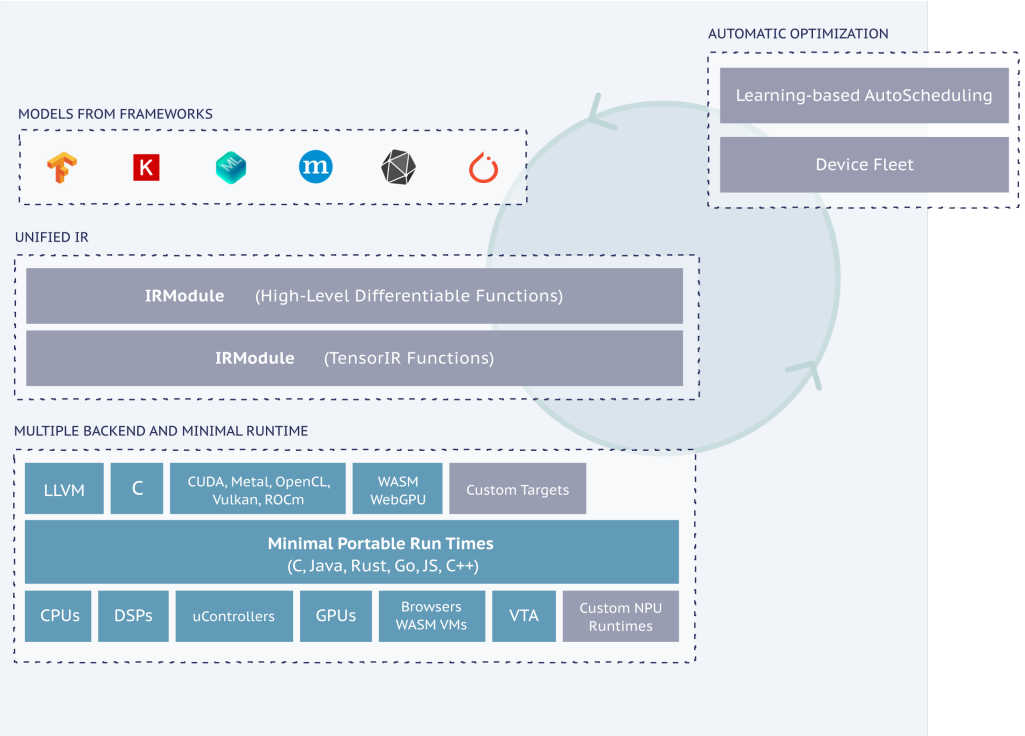
ONNX Runtime

- 微软主要的AI推理解决方案
- 大量用于微软的产品，包括office
- WebGPU后端积极开发中



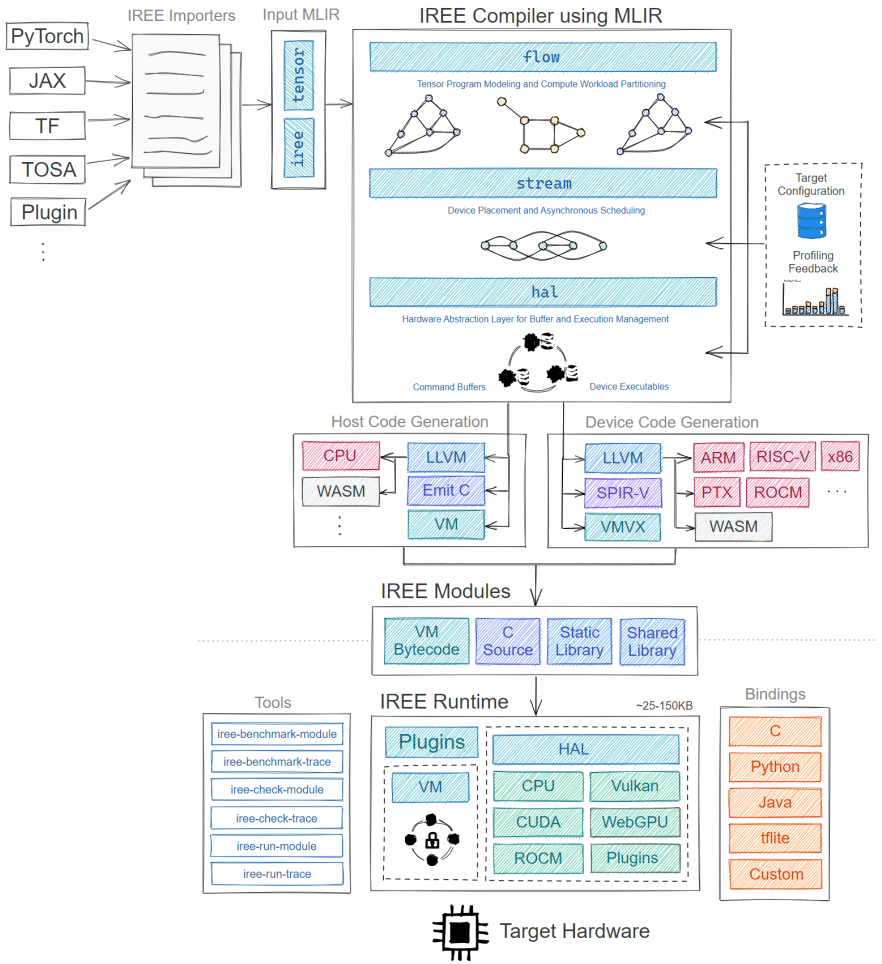
TVM

- 端到端的AI编译器框架
- 广泛的模型支持(PyTorch, ONNX, TensorFlow)
- WebGPU后端性能接近native
- WebGPU后端合作进行中
- 实例
  - WebGPU Stable Diffusion
  - WebGPU LLM (RedPajama-3B, Vicuna-7B)



IREE

- IREE (Intermediate Representation Execution Environment)是基于MLIR (Multi-Level Intermediate Representation)的编译器和运行时
- WebGPU后端开发中

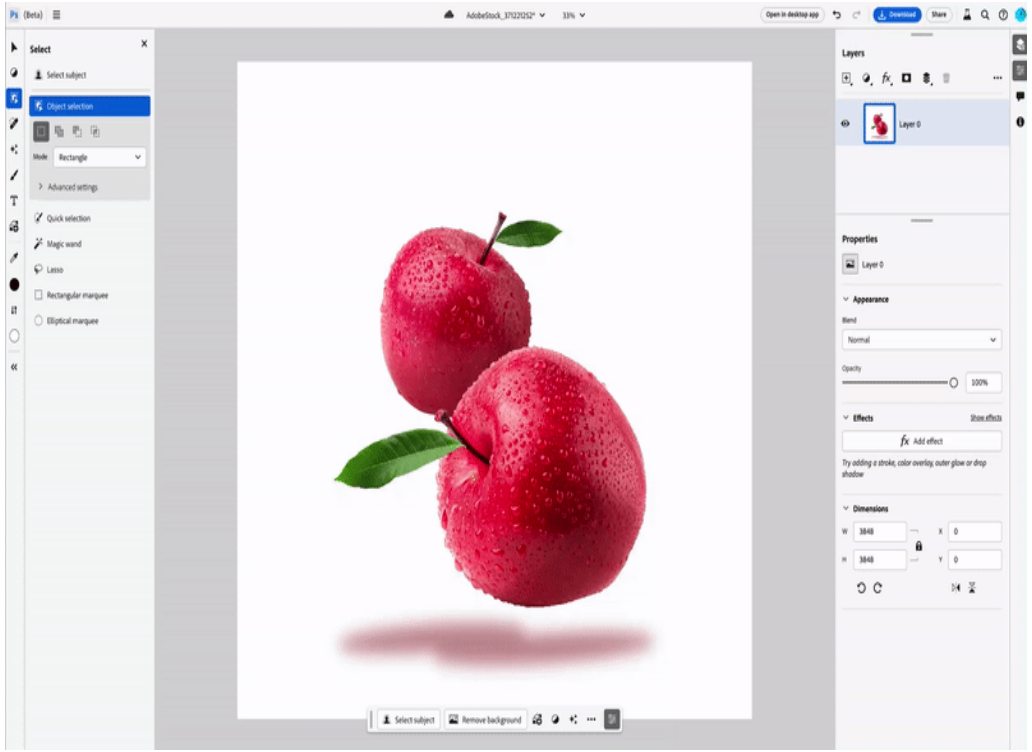


Transformers.js

- Hugging Face transformers Python库的Web版本
- Whisper Demo很快就达到100万点击量
- 作者Joshua Lochner (@xenova, Hugging Face) : “WebGPU support is the next big thing on the todo list”

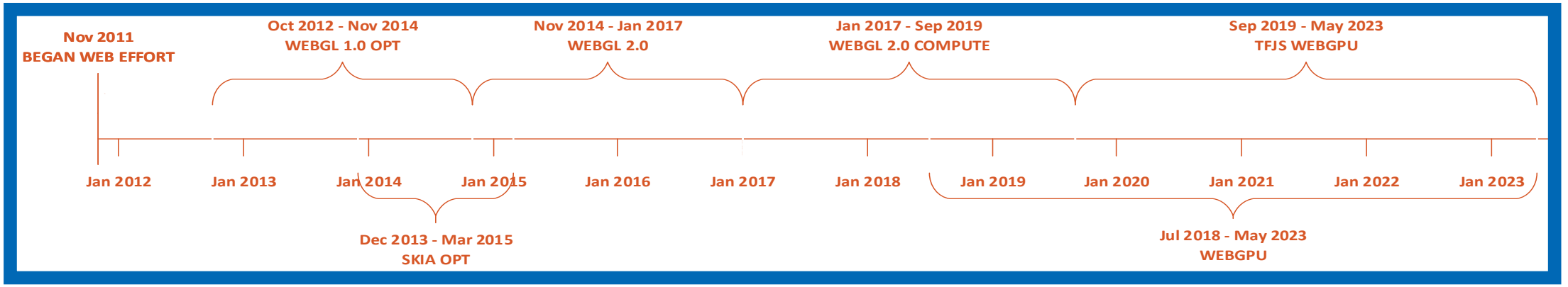
Photoshop Web

- 解决方案基于TensorFlow.js，从而可以方便切换WASM, WebGL, WebGPU后端
- 根据模型大小，灵活切换cloud和on-device方案



英特尔“Web图形和Web AI”团队

英特尔“Web图形和Web AI”团队



- 2011年11月开始Web相关工作
- 贡献了4000+ patch到Web Graphics标准, CTS (Conformance Test Suite), Chromium及相关项目(ANGLE, Dawn等)。在这些项目中担任committer甚至owner角色
- 英特尔在Khronos WebGL Working Group, W3C WebGPU Working Group和W3C Web Machine Learning Working Group的代表
- 英特尔平台所有Web图形问题的负责人

合作交流，携手共进！

- 邮箱: yang.gu@intel.com
- 任何Web图形的问题，欢迎发送到 https://github.com/webatintel/webgraphicsforum/issues