Updates on WebGPU and WebAI

Yang Gu
Manager of Intel WebGraphics and WebAI Team
May 28, 2024
Latest News
Google I/O 2024

WebGPU and Web Assembly are the backbone technologies that enable on-device AI on the web.
Microsoft Build 2024

Enjoy the power of Phi-3 with ONNX Runtime WebGPU
Core Updates
WebGPU Shading Language
W3C Working Draft 14 May 2024

More details about this document
This version:
https://www.w3.org/TR/2024/WD-webgpu-20240514/
Latest published version:
https://www.w3.org/TR/webgpu/
Editor's Draft:
https://github.com/w3c/webgpu
Previous Versions:
https://www.w3.org/TR/2024/WD-webgpu-20240517/
History:
https://www.w3.org/standards/history/webgpu/

TABLE OF CONTENTS
1 Introduction
1.1 Overview
1.2 Syntax Notation
1.3 Mathematical Terms and Notation
2 WGL Module
2.1 Shader Lifecycle
2.2 Errors
2.3 Diagnostics
2.3.1 Diagnostic Processing
2.3.2 Filterable Triggering Rules
2.3.3 Diagnostic Filtering
2.4 Limits
3 Textual Structure
3.1 Parsing
3.2 BlankSpace and Line Breaks
3.3 Comments
3.4 Tokens
3.5 Literals
3.5.1 Boolean Literals
3.5.2 Numeric Literals
3.6 Keywords
3.7 Identifiers
3.7.1 Identifier Compositional
3.8 Context-Dependent Names
3.8.1 Attribute Names
3.8.2 Built-in Value Names
3.8.3 Diagnostic Rule Names
3.8.4 Diagnostic Severity Control Names

---

Milestone 0
No due date  ① Last updated 3 days ago
Formerly “V1.0”. Fixes or spec work for existing V1.0 features.

Milestone 1
No due date  ① Last updated 3 days ago
This is the current milestone for new features that we may land in the spec when they’re resolved. Later milestones mean we’re not landing them yet.

Milestone 2
No due date  ① Last updated 5 days ago

Milestone 3+
No due date  ① Last updated 12 days ago
Items that need to be triaged into “Milestone 3” vs “Milestone 4+” when we define Milestone 3. (The items initially here were formerly “Post-V1.”)
Browser Status

• Overall
  • WebGPU CG was created in Feb 2017. Contributions come from all major browser vendors, Intel and individuals
  • Chrome (MS Edge is the same) supports WebGPU on Windows, ChromeOS and MacOS in M113 (May 2023)
  • Chrome supports WebGPU on Android in M121 (Feb 2024)
  • Safari supports WebGPU in Technical Preview 185
  • Firefox supports WebGPU in Nightly

• Specific features in Chrome
  • DXC (shader compiler) support
  • Timestamp Query, F16, DP4A (INT8), etc.
  • Graphite D3D11 (Replacement of Skia Ganesh)
  • What’s New in WebGPU for more details
Upcoming New Features

- **Push Constants**
- **Wave Matrix**
- **Ray Tracing**
- **Variable Rate Shading**

- **Subgroup**
- **Mesh Shading**
WebGPU Native

- WebGPU is a graphics and compute API for both the web and native

- Cross-platform

- WebGPU headers
Fast Growing Ecosystem
Game Engine -  

- Most popular game engine
- Unity announced the official support of WebGPU in Unity 6
- Demo link and source code
Game Engine –

- Open source HTML5 game engine
- WebGPU support has officially arrived in the PlayCanvas Editor on Apr 18, 2024
• Provides software products and services for the architecture, engineering, construction, manufacturing, media, education, and entertainment industries.

• USD and MaterialX on the WebGPU (Hydra Storm Renderer)
Video Conference - zoom

- Popular video conference solution
- WebGPU for rendering was officially released
AI Framework - TensorFlow.js

• WebAI framework by Google and Intel has been working on its WebGPU backend since Sep 2019

• WebGPU backend was officially released in May 2023
AI Framework – ONNX Runtime

- Microsoft’s major machine learning framework for both native and the web

- Intel joined the WebGPU EP (Execution Provider) effort in July 2023

- 1.17 release on Feb 3 is the first official release of WebGPU EP
AI Framework 🐉 Transformers.js

- Functionally equivalent to HuggingFace’s transformers python library
- Base on ONNX Runtime Web, WASM and WebGPU (V3)
- Maintained by Joshua Lochner (HuggingFace), and more demos can be found at HF
AI Framework – tvm

- An End to End Machine Learning Compiler Framework for CPUs, GPUs and accelerators
- Web solution is based on WebGPU, and already ported many popular LLM models to WebGPU
- Intel contributed F16, while DP4A is WIP
- Enabled Llama3 WebGPU on the day 0 when it was released on Apr 19, 2024
AI Framework - MediaPipe

- **LLM inference**
  - Falcon 1.3B
  - Gemma 2.5B
  - Phi-2 2.7B
  - Stable LM 2.8B

- **Gemini API**
Browser - chrome

• Gemini Nano Integration
  • Integrated into Chrome, and will be released in M126 in June
  • The solution is based on WebGPU
  • “Help me write” feature and High-level APIs

• Graphite Project to replace Skia Ganesh
交流合作，共建生态！

- 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