Cerebras Overview

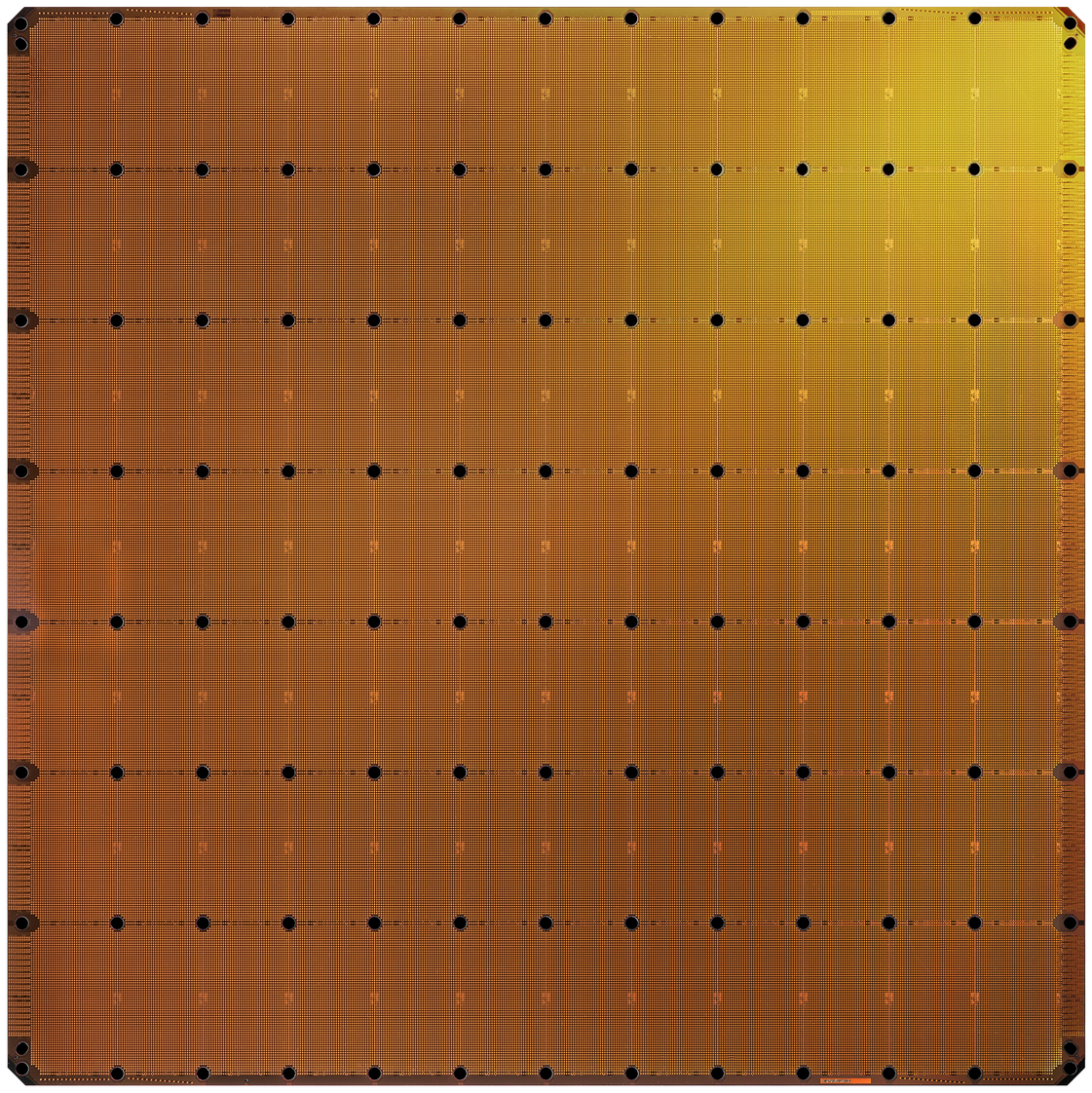
September 2025

Table of Contents

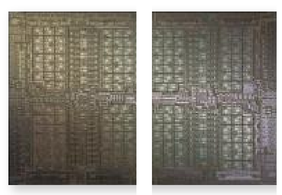
Introduction

Cerebras is a deep tech AI company that delivers the world's fastest AI infrastructure - 20x faster than Nvidia GPU - unlocking use cases that were previously impractical like blazing-fast code gen, instant & intelligent search, life-like virtual assistants, and more.

Cerebras Products



Cerebras
Wafer Scale Engine 3



Nvidia GPU
B200

CLOUD

Serve open models in seconds

Including Llama, DeepSeek, Qwen, and more with an API key

DEDICATED

Scale customer models

On dedicated capacity via a private cloud API/endpoint

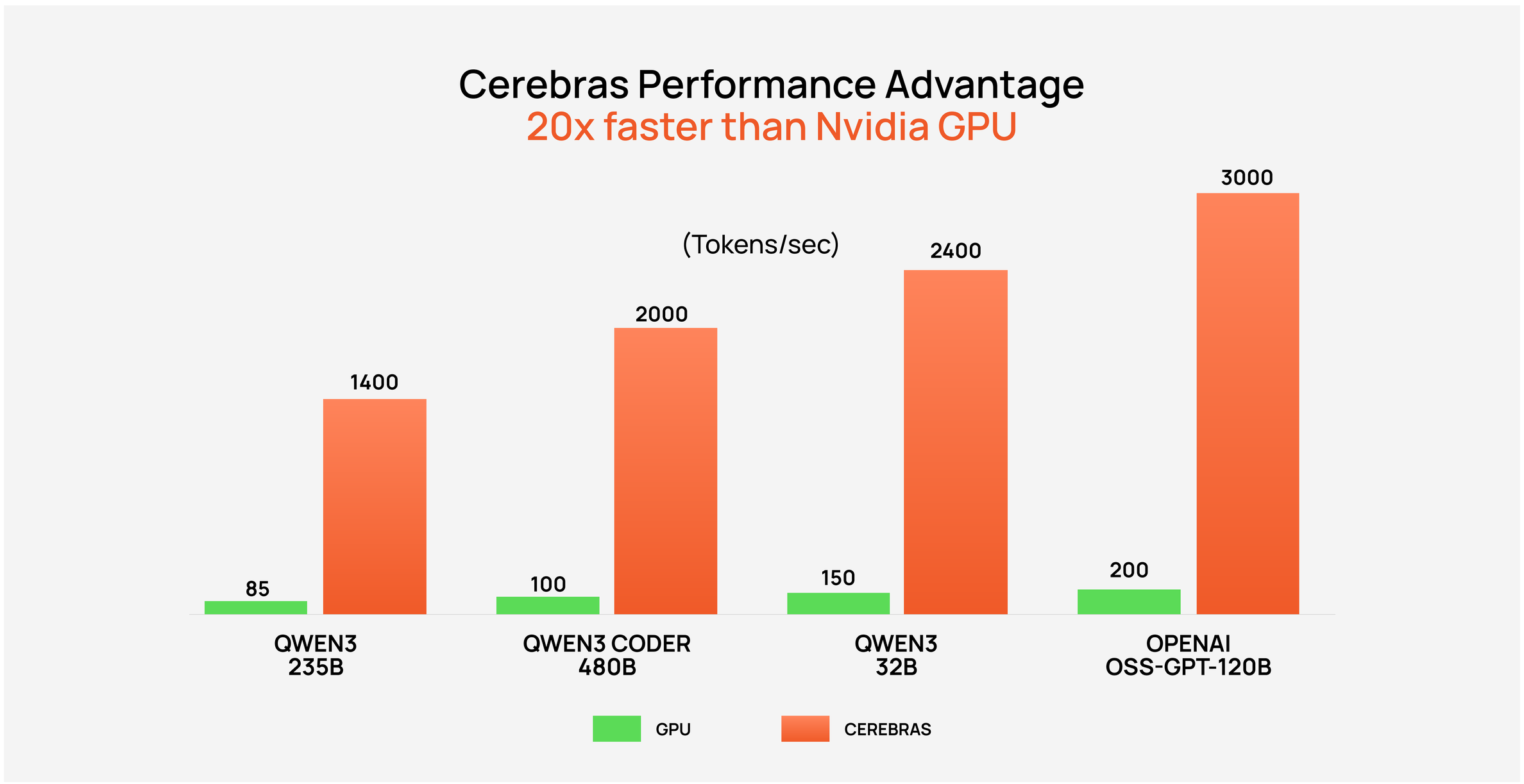
ON-PREM

Deploy on-prem for full control

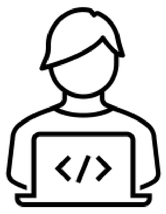
Of models, data and infrastructure in your data center or private cloud

Fastest AI Inference

Cerebras' wafer-scale technology delivers AI speed & latency that no number of GPUs can match .



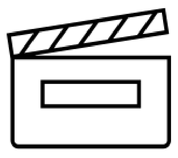
Customer Use Cases



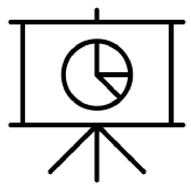
Blazing-Fast
Code Gen

Instant, Intelligent
Search

Life-Like
Virtual Assistants



Lightning Quick
Content Creation



Real-Time
Data Analysis

And More New
Use Cases

Cerebras Powers the World's Leading AI Companies

Meta
perplexity
Mistral AI
GSK
Hugging Face
MAYO CLINIC
IBM
AlphaSense
core42
ALEPH ALPHA
OpenRouter
Quora
C3.ai
TRILOGY
Sandia National Laboratories
Argonne NATIONAL LABORATORY
Lawrence Livermore National Laboratory
THE UNIVERSITY OF EDINBURGH
TotalEnergies
cantina
era
Delphi
ninjatech.ai
PSC
norby
PLURAL·AI
TAKO
tavus
interface.ai
Podium
callin.io
certainly health
personal.ai
Capitol AI
UNEEQ
Lilac Labs
sintra
vivai
oxle
nlx
Triple Whale
KASTLE

Learn More

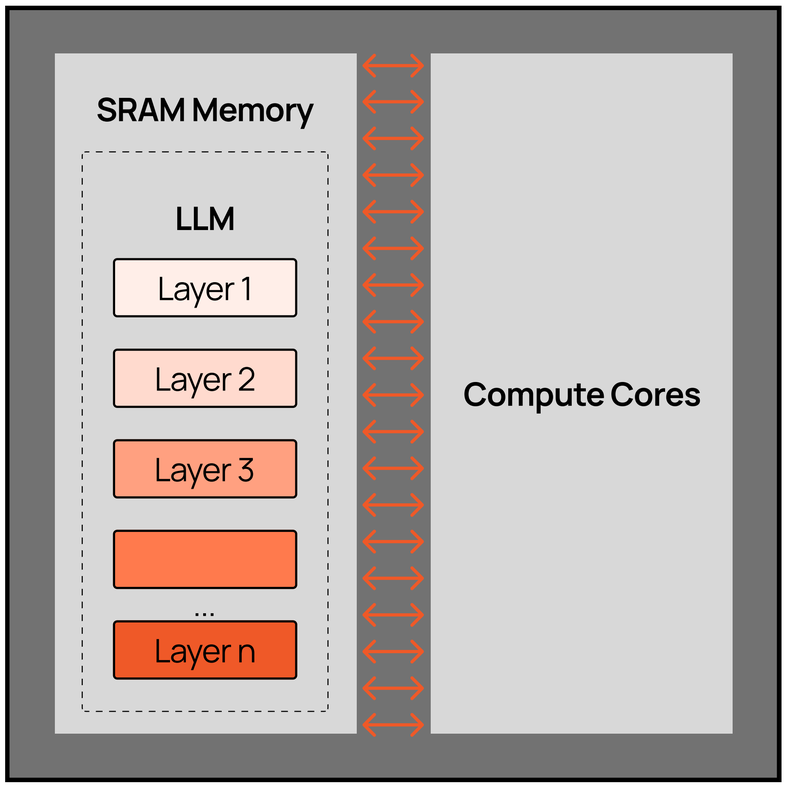
Learn About Cerebras Offerings: https://www.cerebras.ai/

Jumpstart Video: https://www.youtube.com/watch?v=4tbeWbAfxiw

Cerebras Advantage

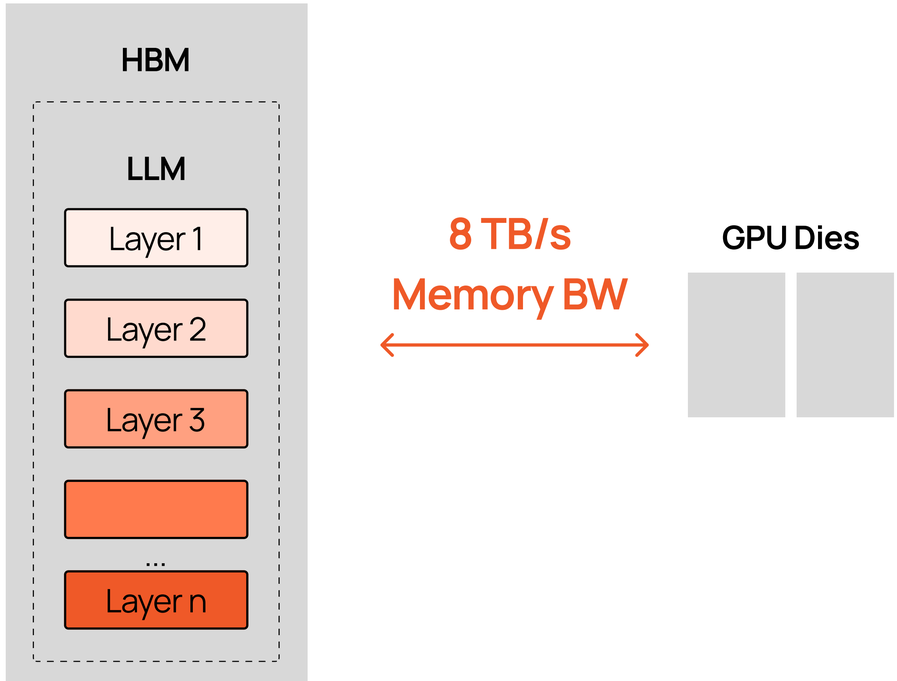
The Wafer-Scale Inference Advantage: massive memory bandwidth and easy scaling delivers blazing fast, efficient inference.

Wafer-Scale Engine



21,00 TB/s
Memory BW

NVIDIA B200



The Cerebras Wafer-Scale Engine is purpose-built for ultra-fast AI. No number of GPUs can match its speed.
Designed for builders who want to do extraordinary things.