Microsoft Bing Speeds Ad Delivery With NVIDIA Triton

Inference software enables shift to NVIDIA A100 Tensor Core GPUs, delivering 7x throughput for the search giant.

Author: Shankar Chandrasekaran

Jiusheng Chen’s team just got accelerated.

They’re delivering personalized ads to users of Microsoft Bing with 7x throughput at reduced cost, thanks to NVIDIA Triton Inference Server running on NVIDIA A100 Tensor Core GPUs .

It’s an amazing achievement for the principal software engineering manager and his crew.

Bing’s ad service uses hundreds of models that are constantly evolving. Each must respond to a request within as little as 10 milliseconds, about 10x faster than the blink of an eye.

The latest speedup got its start with two innovations the team delivered to make AI models run faster: Bang and EL-Attention .

Together, they apply sophisticated techniques to do more work in less time with less computer memory. Model training was based on Azure Machine Learning for efficiency.

Next, the team upgraded the ad service from NVIDIA T4 to A100 GPUs.

The latter’s Multi-Instance GPU (MIG) feature lets users split one GPU into several instances.

Chen’s team maxed out the MIG feature, transforming one physical A100 into seven independent ones. That let the team reap a 7x throughput per GPU with inference response in 10ms.

Triton enabled the shift, in part, because it lets users simultaneously run different runtime software, frameworks and AI modes on isolated instances of a single GPU.

The inference software comes in a software container, so it’s easy to deploy. And open-source Triton — also available with enterprise-grade security and support through NVIDIA AI Enterprise — is backed by a community that makes the software better over time.

Accelerating Bing’s ad system with Triton on A100 GPUs is one example of what Chen likes about his job. He gets to witness breakthroughs with AI.

While the scenarios often change, the team’s goal remains the same — creating a win for its users and advertisers.

Original URL: https://blogs.nvidia.com/blog/2023/06/05/microsoft-bing-triton/