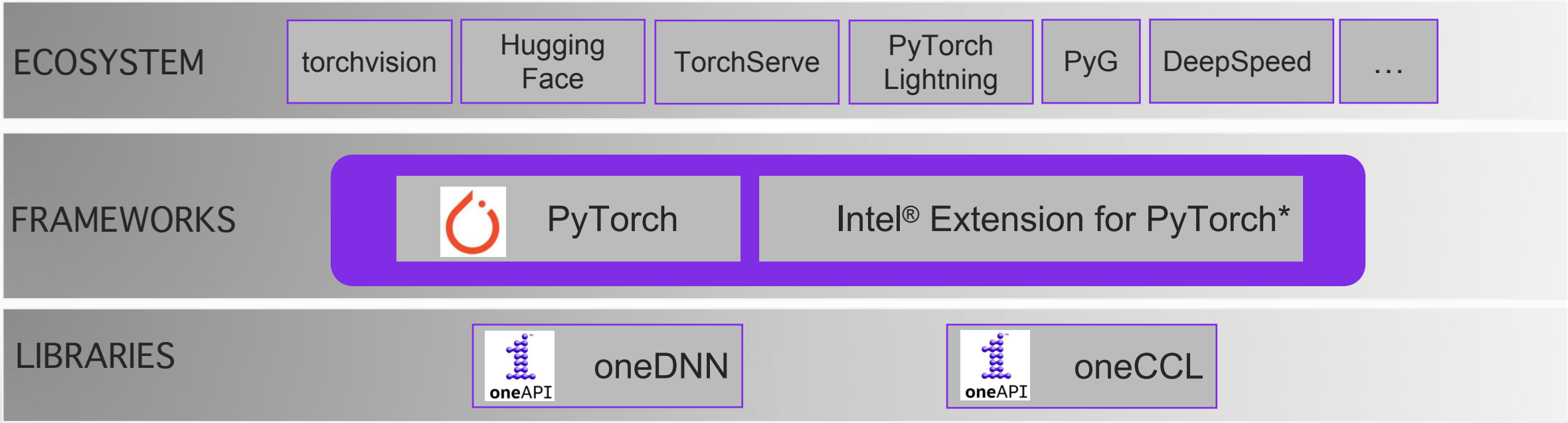
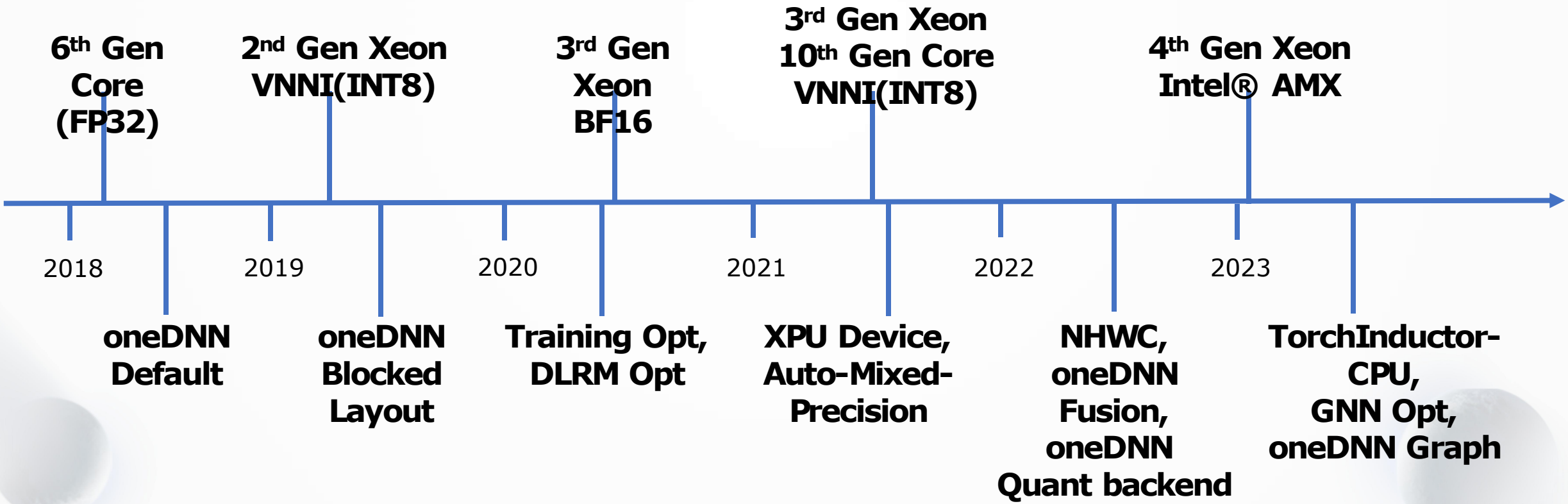

GOTC 2023
全球开源技术峰会
THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE
OPEN SOURCE, INTO THE FUTURE
「AI is Everywhere」专场
英特尔的 PyTorch 之旅：AI 算力提升和开源软件优化
英特尔亚太研发 马鸣飞
2023年5月28日

概述


ECOSYSTEM
torchvision
Hugging Face
TorchServe
PyTorch Lightning
PyG
DeepSpeed
…
FRAMEWORKS
PyTorch
Intel® Extension for PyTorch*
LIBRARIES
oneAPI
oneDNN
oneAPI
oneCCL
intel
XEON
inside

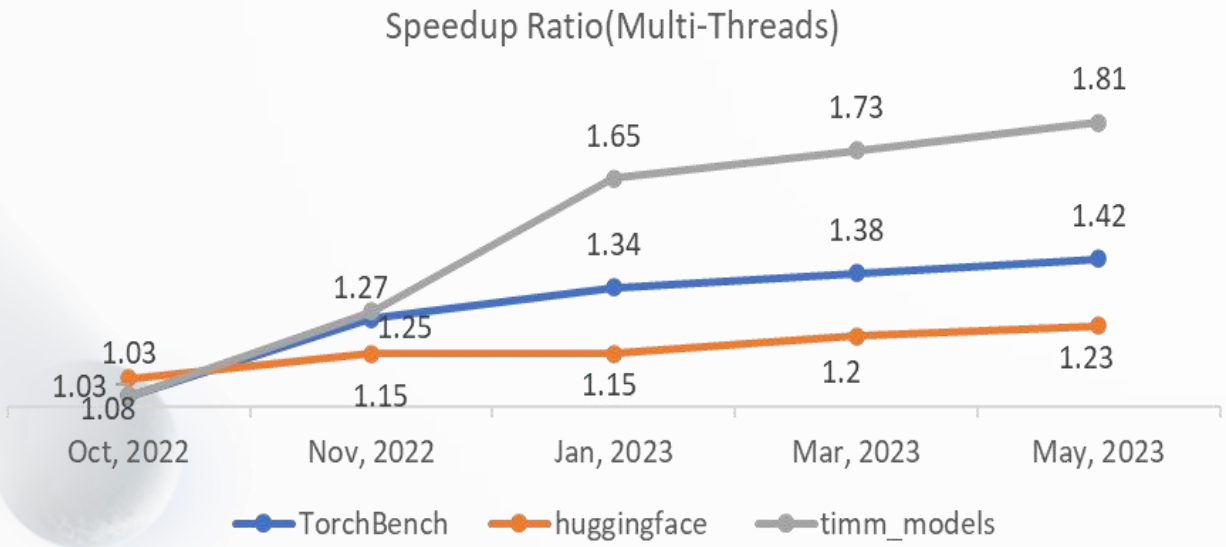
PyTorch Upstream 里程碑

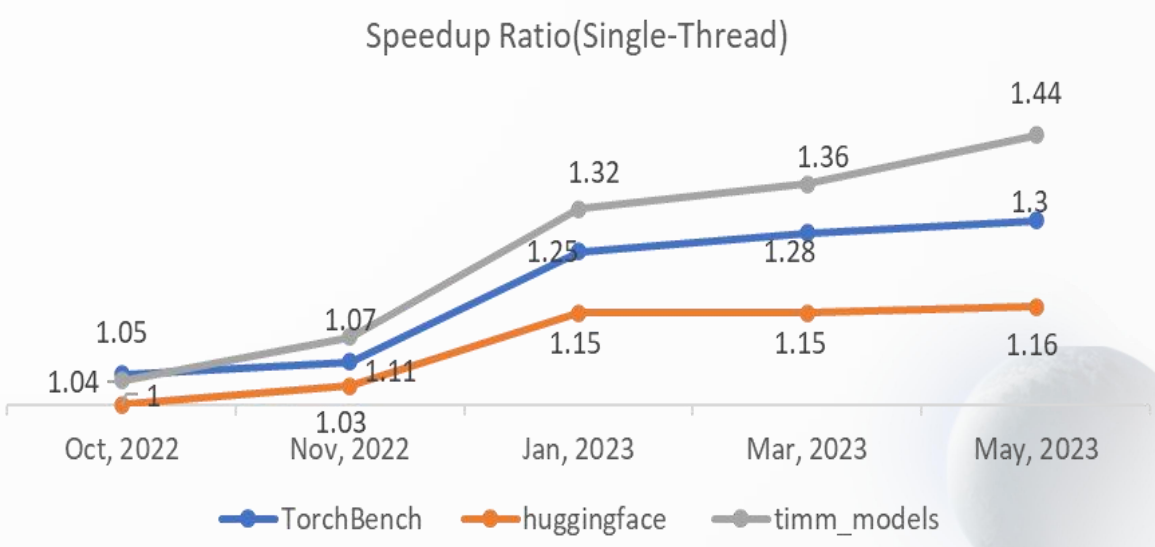


PyTorch 2.0: torch.compile

TorchInductor CPU 后端的相关优化

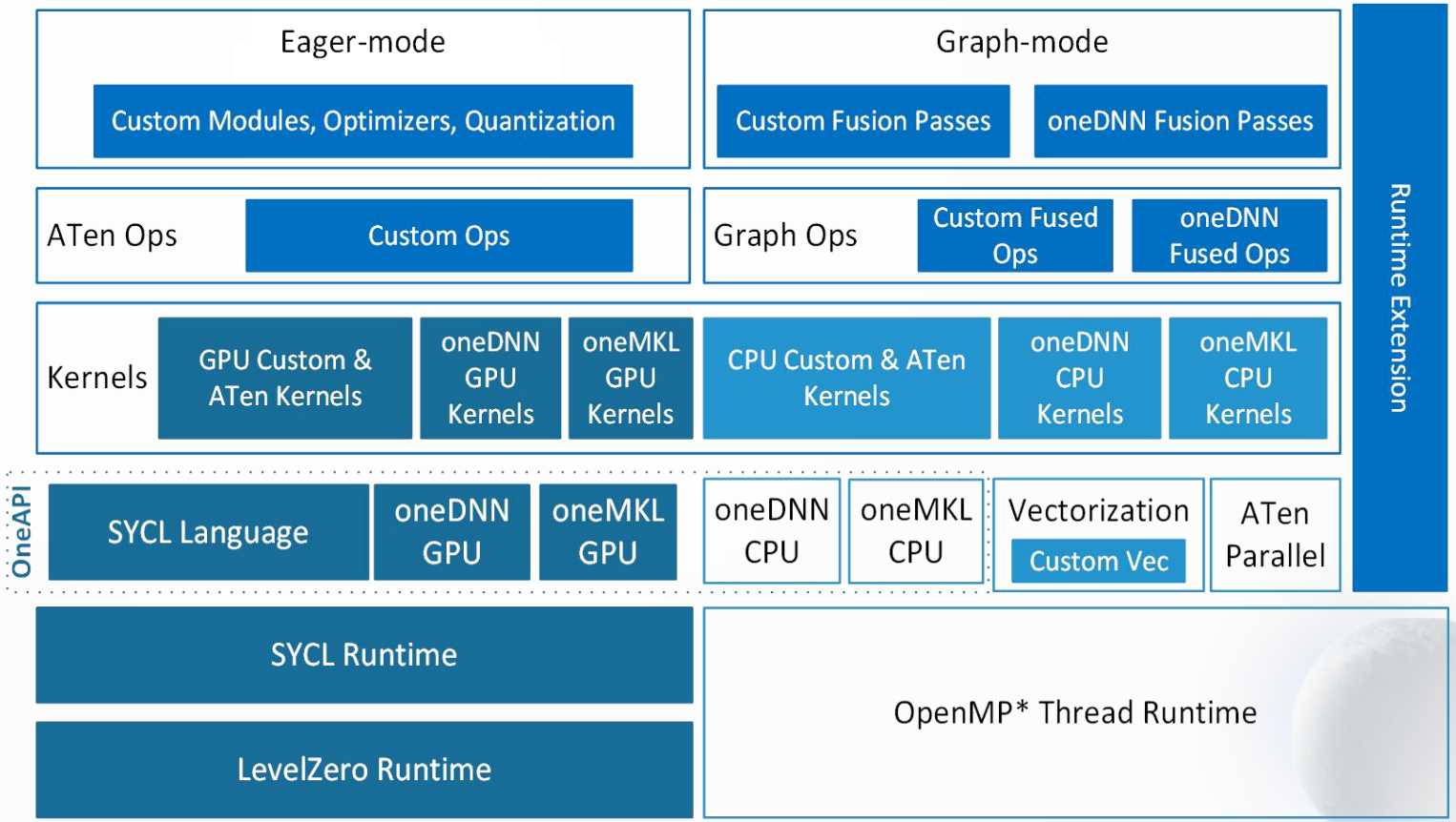
- FP32 部署的优化已在 PyTorch 2.0 上发布
- BF16 和训练相关的优化将在 PyTorch 2.1 上发布





Intel® Extension for PyTorch*

- 一行转换模型：提供格外性能提升
- 适配 Intel 最新硬件 CPU，GPU
- Python，C++ 双接口方便训练+部署



Intel Optimizations for PyTorch

Operator

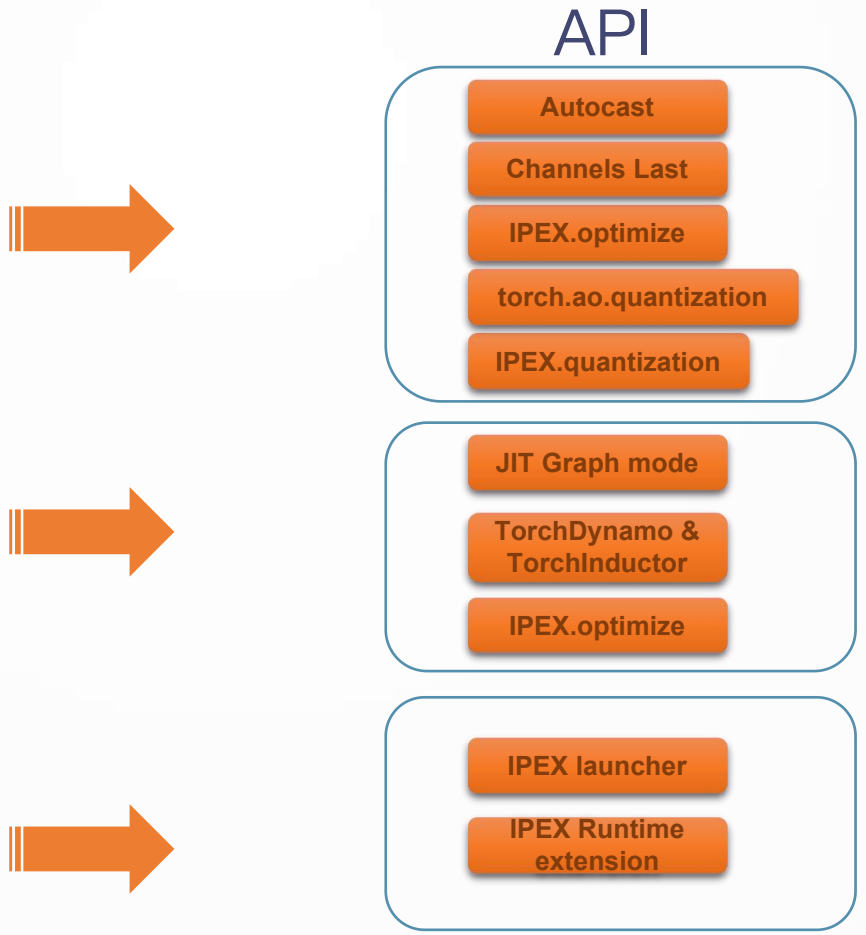
- Vectorization and Parallelization
- Low-precision computation BF16/INT8
- Auto-Mixed-Precision (AMP)
- Data layout optimization

Graph

- Constant folding to reduce compute
- Op fusion for better cache locality

Runtime

- Thread affinization and multi-streams
- Memory buffer pooling
- GPU runtime
- Launcher



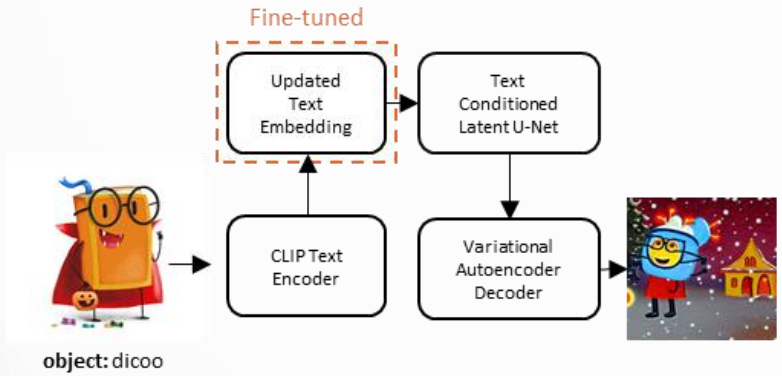
生成式AI: Stable Diffusion

Create Your Own Stable Diffusion

Few-shot Fine-tuning

5 minutes

4 SPR nodes



Accelerated Stable Diffusion Inference

Low-precision Inference

5 seconds

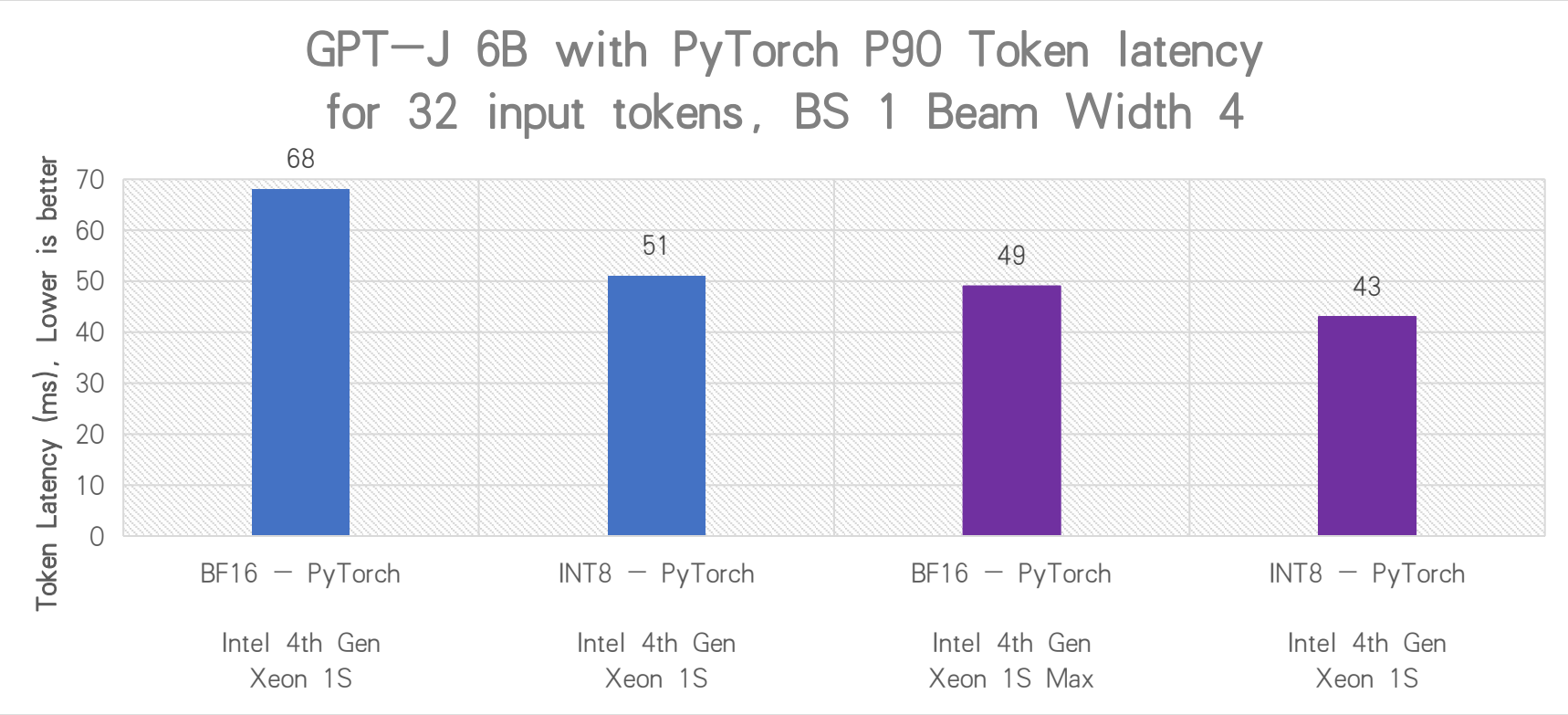
1 SPR chip

Optimizations upstreamed to Hugging Face Diffusers and Optimum-Intel

Demo on Intel Dev Cloud: https://stablediffusion.eglb.intel.com/
Demo on AWS: https://huggingface.co/spaces/Intel/Stable-Diffusion

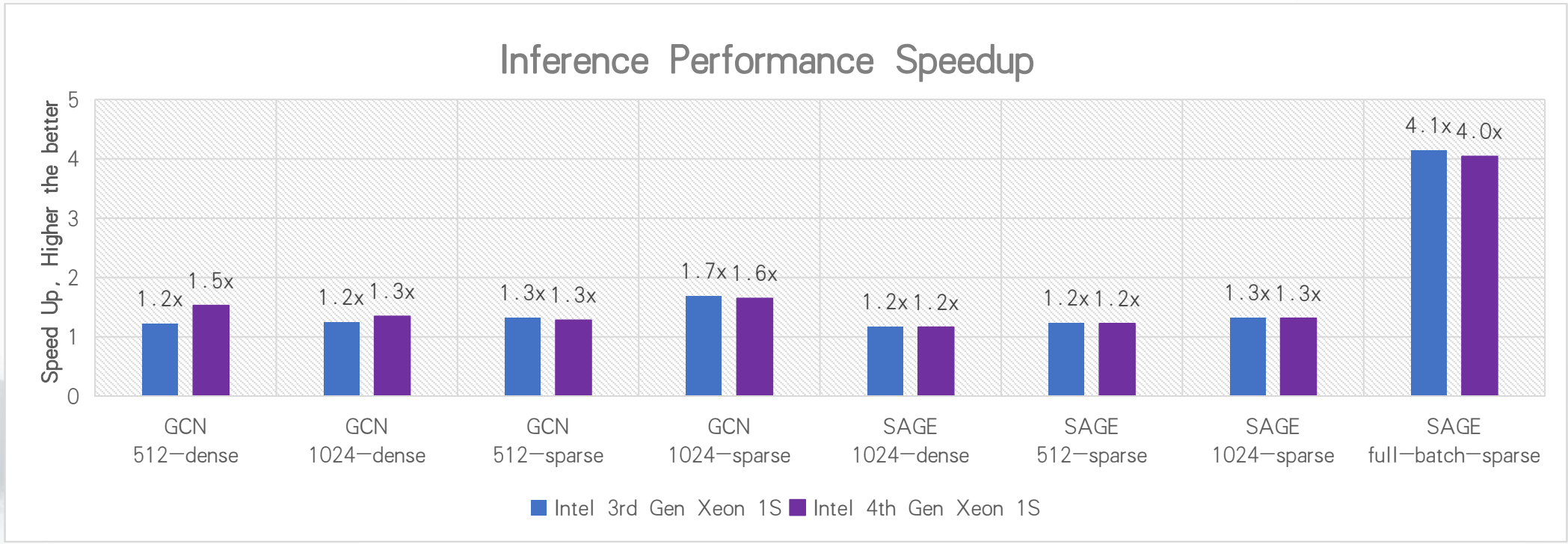
大语言模型: LLM

- Intel® Extension for PyTorch* 实现针对 MHA 和 FFN 的模块的 Kernel Fusion
- 应用低精度 BF16, INT8，减少内存开销



图神经网络： PyG

- PyG - 建立在 PyTorch 基础上的库，用于编写和训练图神经网络（GNN)
- PyTorch 2.0 及 PyG 2.3 中的相关优化提供至多 4x 的性能提升



https://www.pyg.org/ns-newsarticle-accelerating-pyg-on-intel-cpus

PyTorch + Intel GPU

- Intel 独立 GPU 上提供完整的 PyTorch 支持，如 Auto Mixed Precision (AMP), Channels Last, DPC++ Extension, graph fusion 等等。
- 通过注册 “XPU”设备，实现标准的 PyTorch 前端编程模型。
- 后端 Kernel 通过 oneAPI 编程模型实现。
- Intel GPU 支持随 Intel Extension for PyTorch 发布，目前支持 torch 1.13。
- https://www.intel.com/content/www/us/en/developer/articles/technical/introducing-intel-extension-for-pytorch-for-gpus.html

GOTC

THANKS

全球开源技术峰会

THE GLOBAL OPENSOURCE TECHNOLOGY CONFERENCE