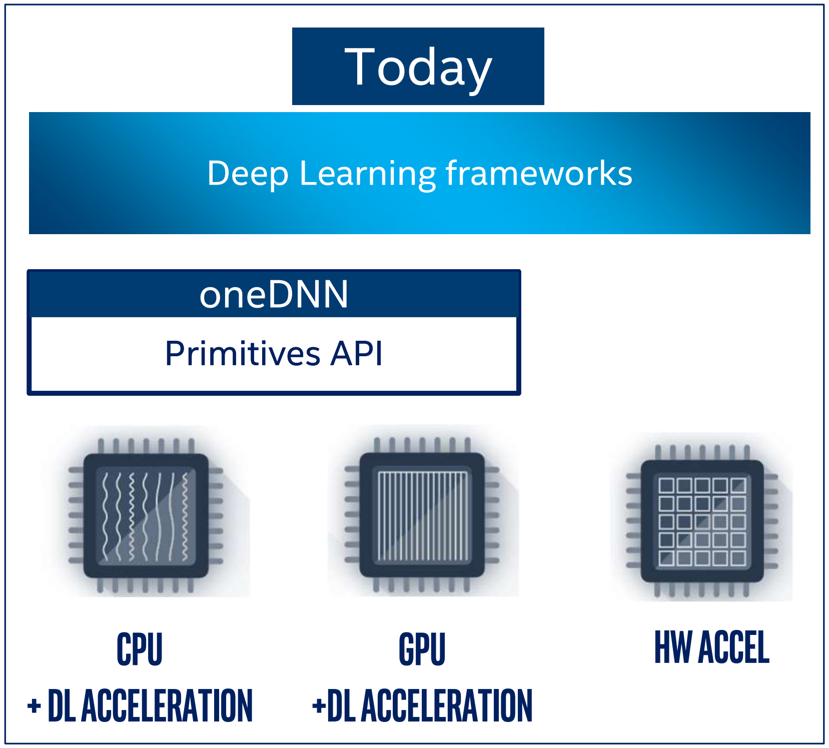
oneDNN Graph API

Jianhui Li
Principal Engineer, Intel

oneDNN is evolving…



Today

Deep Learning frameworks

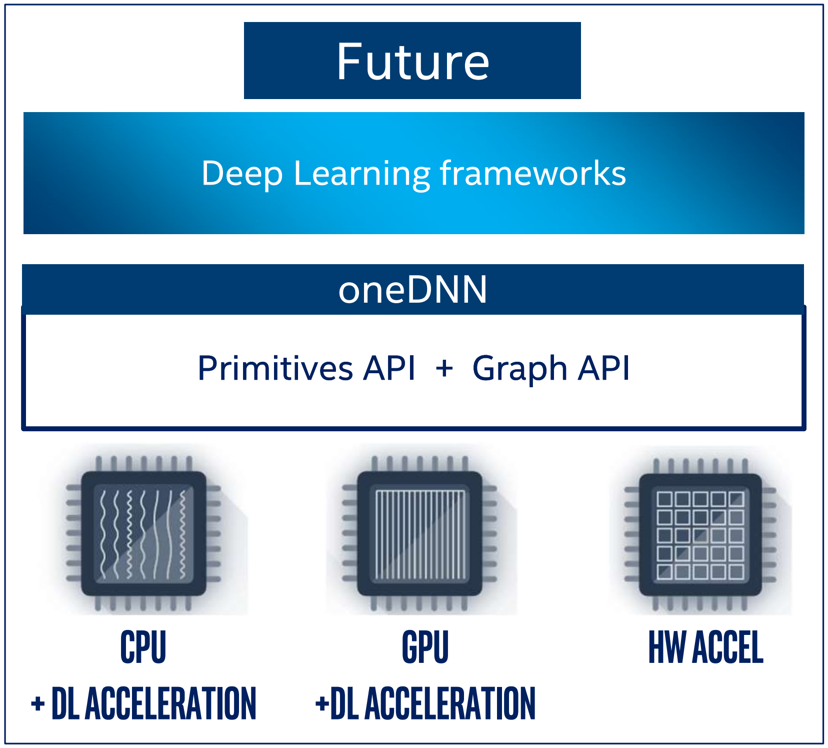
oneDNN

Primitives API

CPU
+ DL ACCELERATION

GPU
+DL ACCELERATION

HW ACCEL



Future

Deep Learning frameworks

oneDNN

Primitives API + Graph API

CPU
+ DL ACCELERATION

GPU
+DL ACCELERATION

HW ACCEL

- Graph API allows HW backend to maximize performance
- Same integration for multiple AI HW: CPU, GPU, and accelerators

Latest Update from oneDNN Graph

1. SPEC v0.2 preview available on oneAPI SPEC website

https://spec.oneapi.com/onednn-graph/latest/

2. oneDNN Graph API code preview branch on oneDNN github

https://github.com/oneapi-src/oneDNN/tree/dev-graph

3. Pytorch experimental PR available and received positive feedback from FB
https://github.com/pytorch/pytorch/issues/49444

4. TensorFlow experimental PR ready for feedback

https://github.com/Intel-tensorflow/tensorflow/tree/dev-graph/third_party/oneDNNGraph

oneDNN Graph SPEC Roadmap

SPEC0.2	SPEC0.5	SPEC0.8	SPEC1.0	SPEC2.0 and beyond
• oneDNN Graph programing model • FP32/FP16/BF16 Ops for Inference and Training	• Blocked Layout • In-place Support	• Int8 Inference	• V1.0 Finalize	• Control Flow • Dynamic Shape • Custom OP Registration
Q4’20	Q1’21	Q3’21	Q1’22	Future

oneAPI

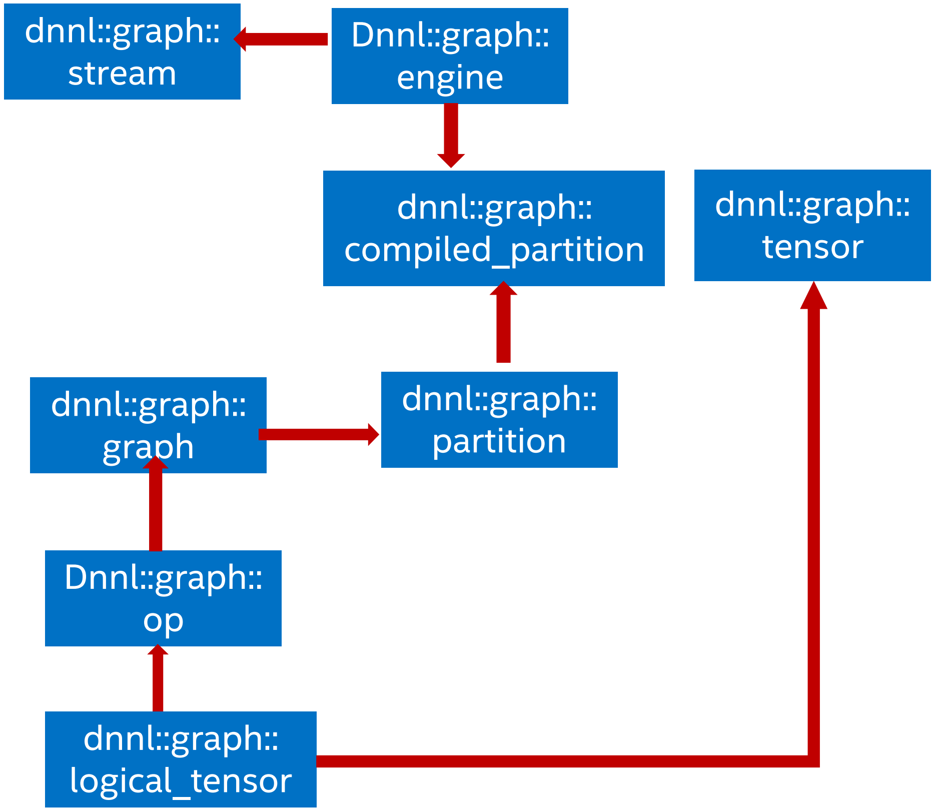
oneDNN Graph programming model

Partition

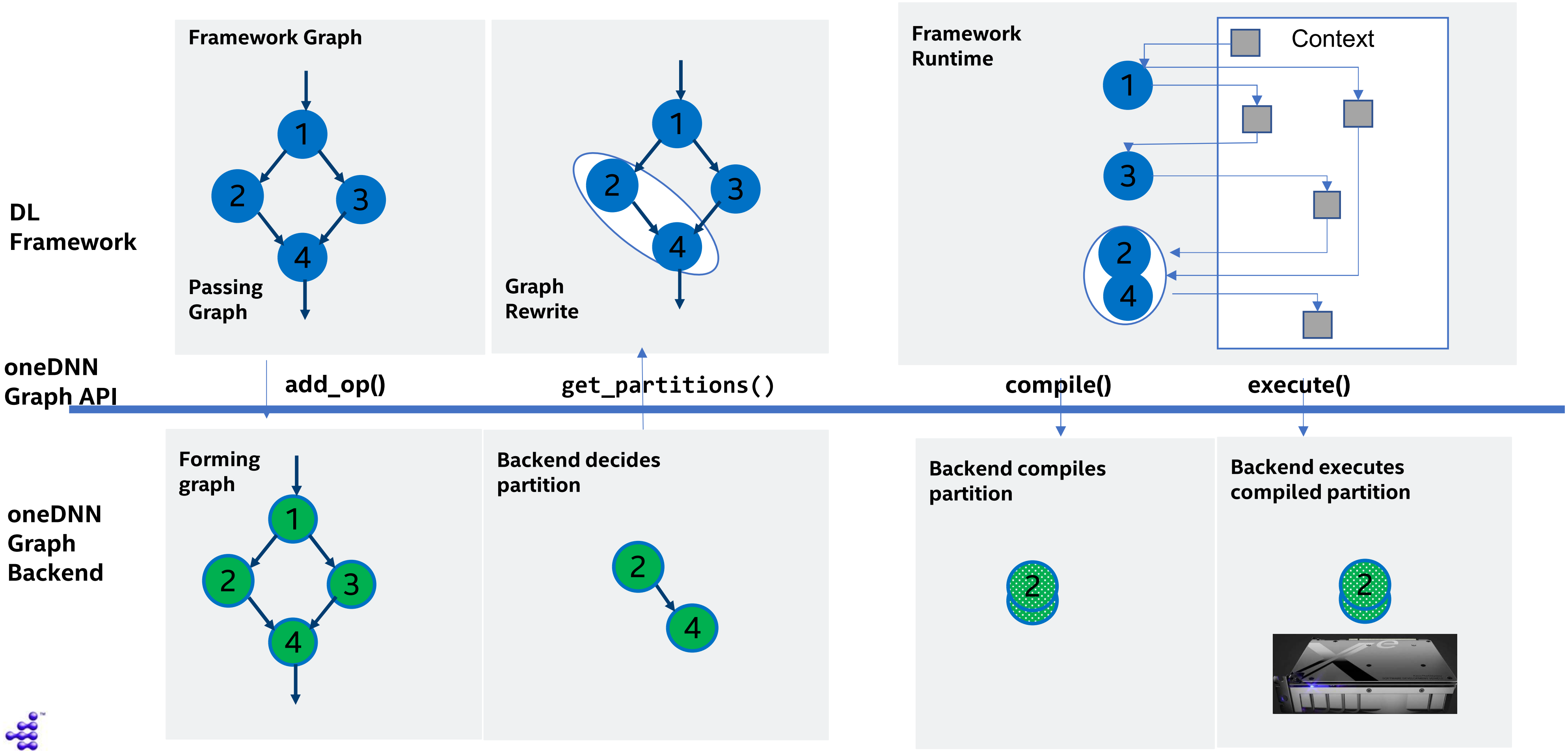
- Logical tensor: tensor's metadata like dims, data type, layout
- Op: DNN op with attributes, associated with input/output logical tensors
- Graph: a collection of Op and logical tensors
- Partition: a subgraph for target specific optimization

Compilation & Execution

- Engine – execution device
- Stream – execution context
- Compiled partition: compiled object for partition
- Tensor: data storage + metadata



oneDNN Graph API



DL Framework

Framework Graph

Passing Graph

Graph Rewrite

Framework Runtime

Context

oneDNN Graph API

add_op()

get_partitions()

compile()

execute()

oneDNN Graph Backend

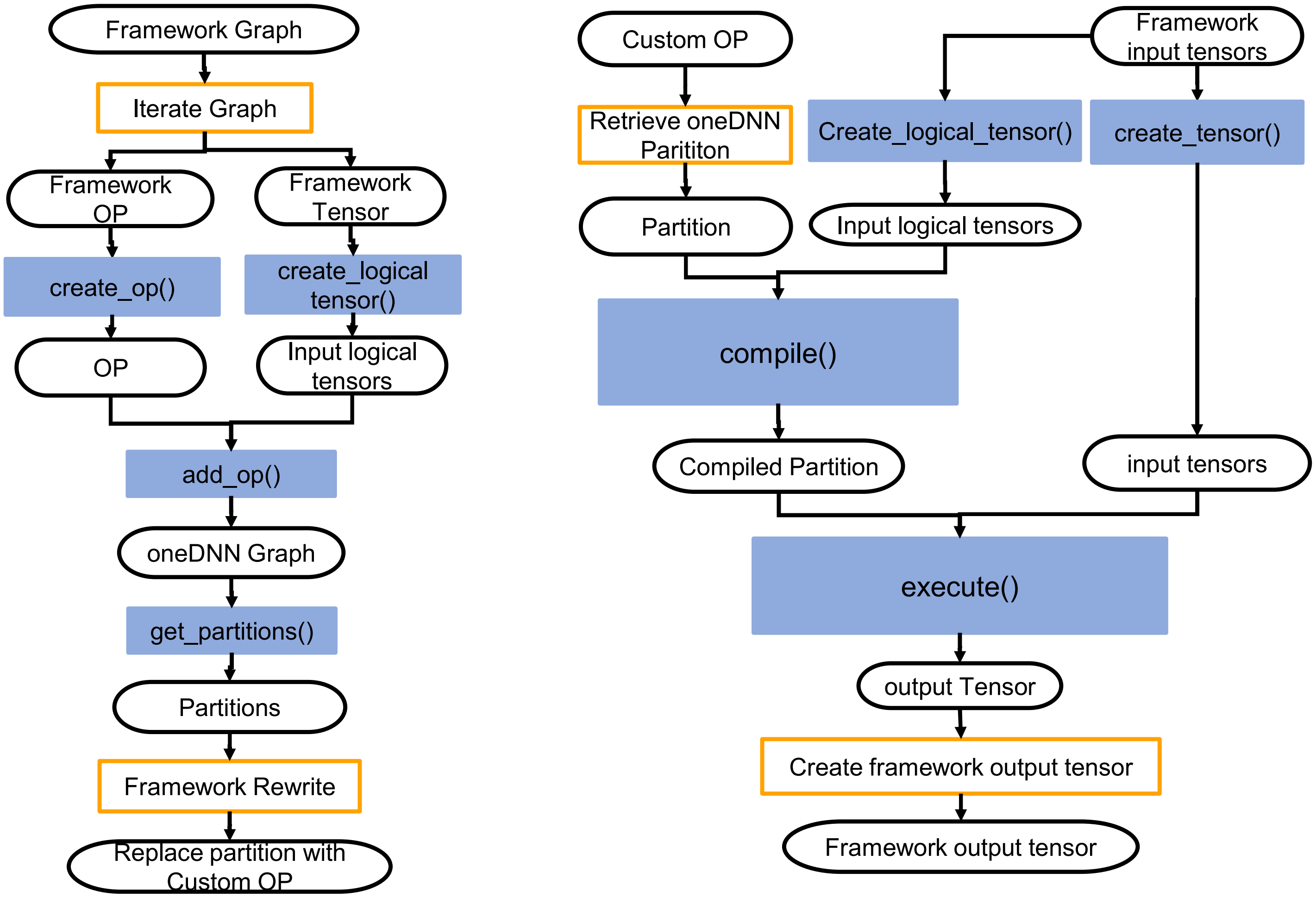
Forming graph

Backend decides partition

Backend compiles partition

Backend executes compiled partition

Framework Integration Flow Graph



Framework Graph

Iterate Graph

Framework OP

Framework Tensor

create_op()

create_logical tensor()

OP

Input logical tensors

add_op()

oneDNN Graph

get_partitions()

Partitions

Framework Rewrite

Replace partition with Custom OP

Custom OP

Retrieve oneDNN Parititon

Partition

Framework input tensors

Create_logical_tensor()

create_tensor()

Input logical tensors

compile()

Compiled Partition

input tensors

execute()

output Tensor

Create framework output tensor

Framework output tensor

input tensors

Framework Object

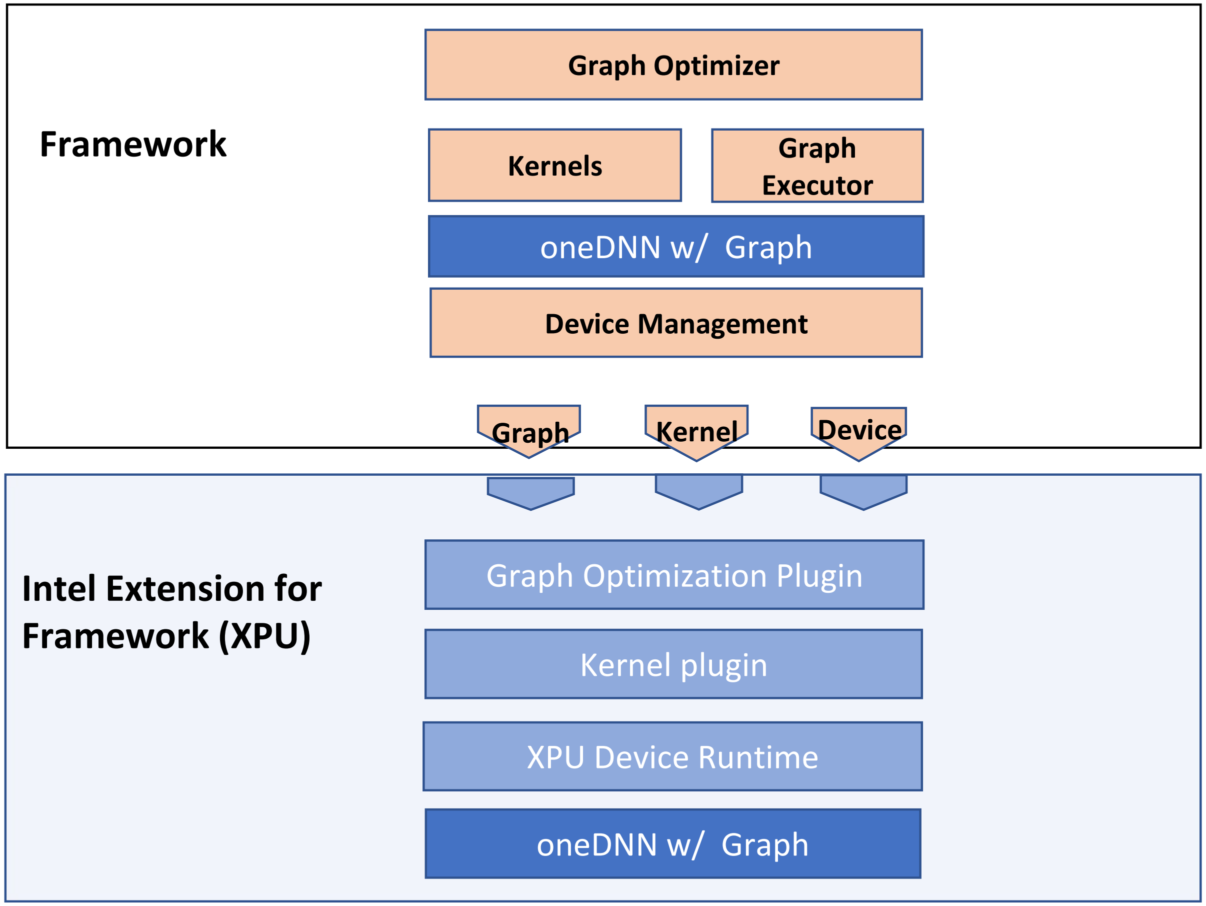
Iterate Graph

Framework Function

create_ten sor()

Framework Function using oneDNN Graph

Framework Integration Scenario



Flexible fusion
Ease of integration

Advanced optimization
Support multiple AI HW

oneAPI™

Thank You!

http://oneapi.com