昇腾CANN系列教程

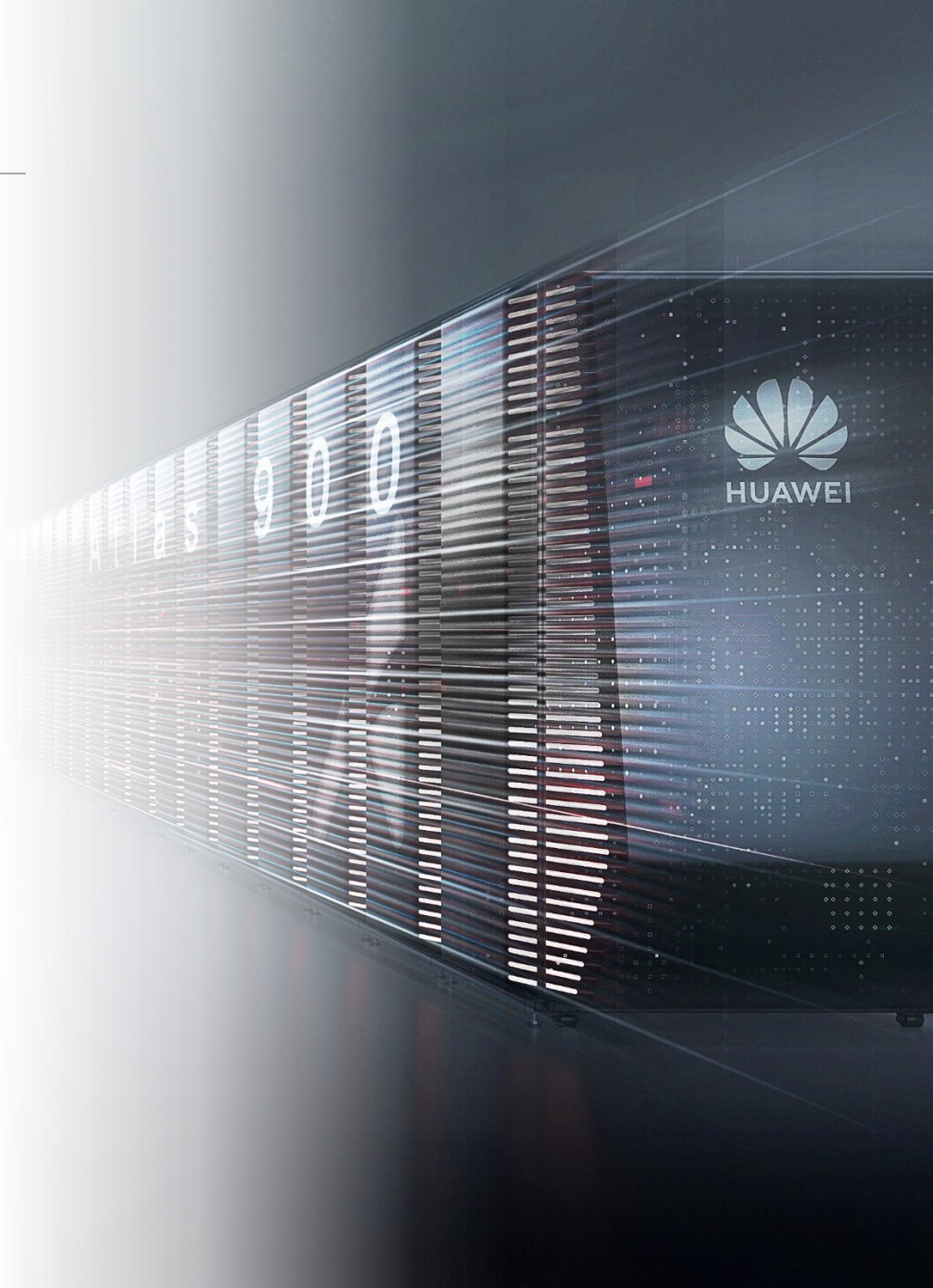
——TBE实战算子开发-DSL方式

HUAWEI

培训目标

- 学完本课程后，您应该能：
  - 运用TBE框架的DSL方式开发自定义算子。
  - 理解TBE算子编译过程。

目录 TABLE OF CONTENT



TBE算子——Hello World

- 目标：

  使用DSL语言实现平方根功能的TBE算子。

- 接口命名：

  sqrt()

- 算子分析：

  Sqrt算子功能是对Tensor中每个原子值求开方，数学表达式子为$\mathbf{y}=\sqrt{x}$。

根据当前TBE框架可支持的计算描述API，采用如下公式来表达Sqrt算子的计算过程：$\mathbf{y = exp(0.5*log(x))}$

Hello World（续）

```
def sqrt(x, y, kernel_name="sqrt"):
    data = tvm.placeholder(x.get( “shape” ), name="data", dtype=x.get( “dtype” ))
    log_val = dsl.vlog(data)
    const_val = tvm.const(0.5, "float32")
    mul_val = dsl.vmuls(log_val, const_val)
    res = dsl.vexp(mul_val)
    with tvm.target.cce():
        sch = dsl.auto_schedule(res)
    config = {"name": kernel_name, "tensor_list": [data, res]}
    dsl.biuld(sch, config)
```

TBE算子代码结构——TBE算子入参

- TBE算子基本入参

  def **sqrt**(x, y, kernel_name="sqrt"):

  - x：输入张量，json格式，属性说明：

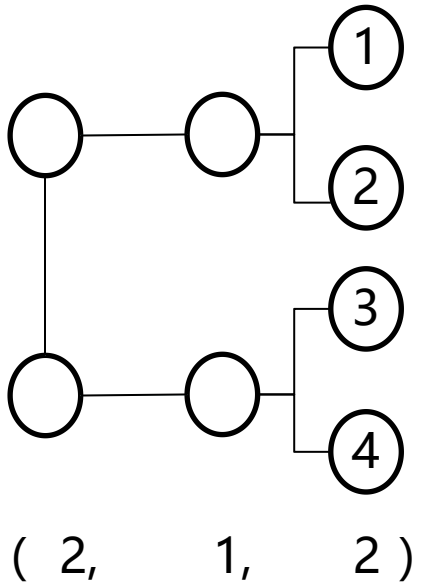
    >shape : Tensor的属性，表示Tensor的形状，用list或tuple类型表示。

    例如（3, 2, 3）、（4, 10）。

    >dtype ： Tensor的数据类型，用字符串类型表示。

    例如“float32”、“float16”、“int8”等。

  - Y:输出张量，json格式，属性同x



TBE算子代码结构——输入占位符

- TBE算子输入占位符

  样例：`data = tvm.placeholder(shape, name="data", dtype=dtype)`

tvm.placeholder()是tvm框架的API，用来为算子执行时接收的数据占位，通俗理解与C语言中%d、%s一样，返回的是一个Tensor对象，上例中使用data表示；入参为shape，name，dtype,是为Tensor对象的属性。

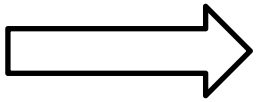
这里的输入是指算子执行时的输入数据，与编译时期入参不同，编译时期入参(x,y)是为了得到算子执行文件的入参。

TBE算子代码结构——定义计算过程

- TBE算子中定义计算过程

  定义计算过程是指使用DSL语言，根据数学算式，描述出实现算子功能的计算步骤，以算子sqrt为例：

$$y = \exp(0.5*\log(x))$$



```
log_val = dsl.vlog(data)
const_val = tvm.const(0.5, "float32")
mul_val = dsl.vmuls(log_val, const_val)
res = dsl.vexp(mul_val)
```

根据sqrt算子的函数表达式，描述sqrt的计算过程如上所示，其中dsl.name()皆为TBE框架的API.

TBE算子代码结构——调度

- TBE算子中调度（schedule）操作

  样例：

```
with tvm.target.cce():
    sch = dsl.auto_schedule(res)
```

计算过程描述完之后，就会做调度；调度是与硬件相关的，功能主要是调整计算过程的逻辑，意图优化计算过程，使计算过程更高效，以及保证计算过程中占用硬件存储空间不会超过上限。

TBE算子代码结构——构建

- TBE算子中的构建操作

  样例： dsl.build(sch, config)

  TBE框架提供了build()API，传入schedule以及相关的配置项，即可完成编译，生成最终硬件可执行文件。

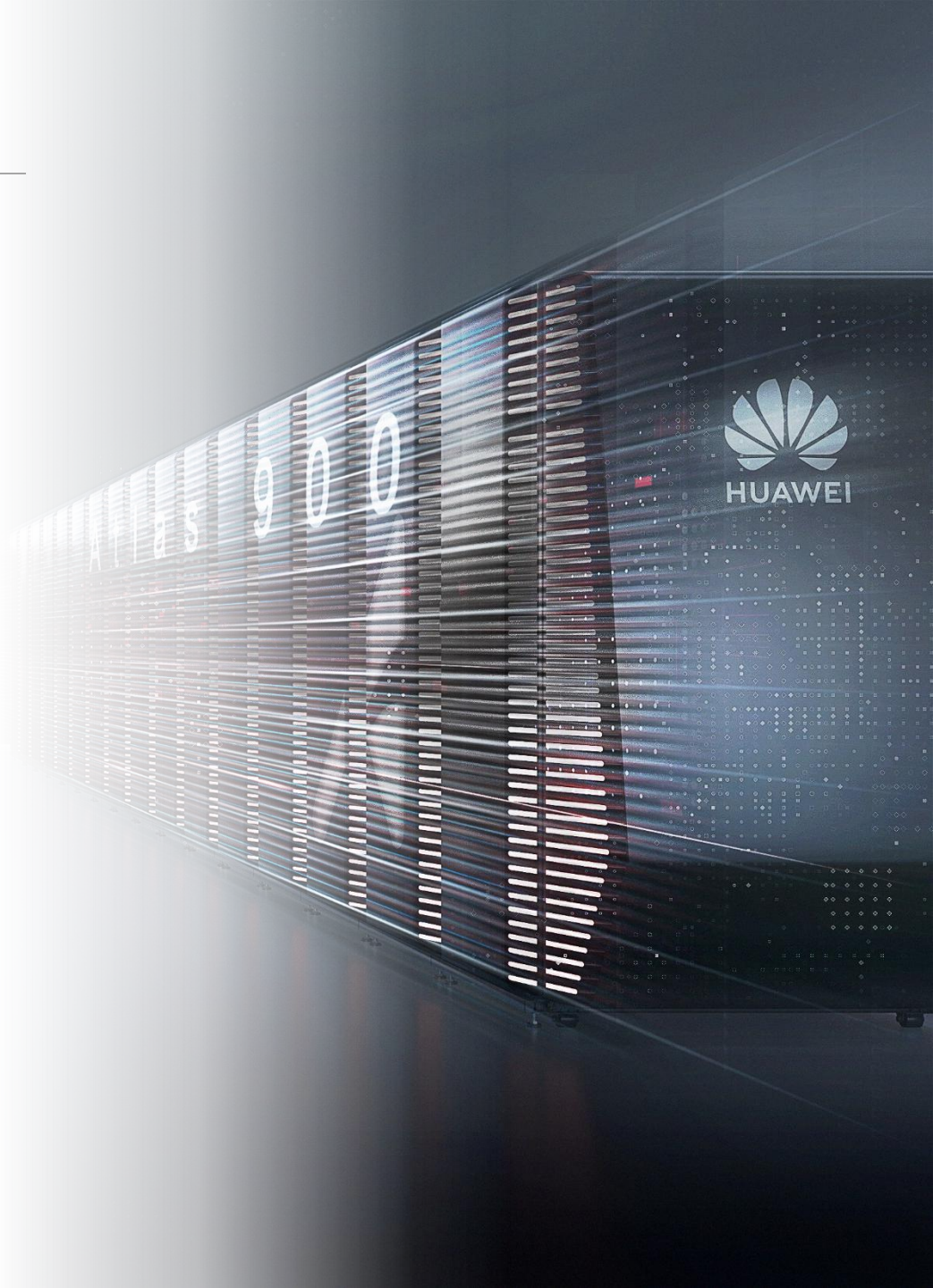
TBE框架ComputeAPI

- TBE框架提供描述计算过程的API：

TBE算子都是调用框架提供的computeAPI来描述计算过程，接口皆为dsl.name的形式；compute API 现根据功能类型可分为以下几类：

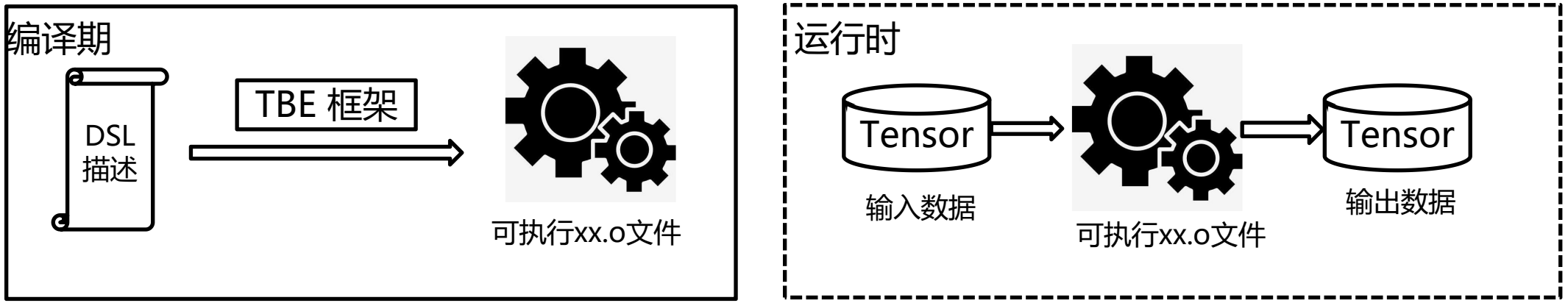
接口分类	简介
Math	对Tensor中每个原子值分别做相同操作的计算接口。
NN	神经网络相关计算接口。
Cast	取整计算接口，对输入Tensor中的每个元素按照一定的规则进行取整操作。
Inplace	对Tensor进行按行相关计算。
Reduce	对Tensor按轴进行相关操作的计算接口。
Matmul	矩阵乘计算。
Gemm	通用矩阵乘计算接口。
卷积	包含2D卷积运算和3D卷积运算的相关接口。
Pooling2D/3D	**2D/3D**池化接口。
Array	在指定轴上对输入Tensor进行重新连接或者切分的接口。

目录 TABLE OF CONTENT



TBE算子编译过程

TBE算子编译与应用



TBE算子经过编译，生成能够在硬件上运行的xxx.o形式的可执行文件，此为TBE算子在TBE框架下执行后的输出件。

TBE算子编译过程

TBE算子编译过程分为DSL->Schedule->pass->codegen四步

- Schedule

  经过Schedule，计算过程描述逻辑发生了转变，针对硬件的存储上限做了切分、合并等操作。

- Pass

  Pass阶段会对计算过程描述进行指令替换，将数学方式表示的计算描述，映射为硬件可以读懂的指令，且会对指令进行优化，以获得更高效的性能，经过pass后，计算过程变为IR表示。

- Codegen

  Codegen是TBE框架执行的最后一步，将Pass产生的IR构建为cce代码，进而经过compile生成二进制的可执行文件。

Thank you.

昇腾开发者社区

http://hiascend.com

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.