SIPU Tile Core常用指令归纳

1 常用指令

1.1 load

数据加载指令，一般用于将数据从global memory或shared memory加载到寄存器，主要支持功能有：

（1）数据形状：一维线性（linear），二维块状（blk）

（2）数据位置：global memory（global），shared memory（shared）

（3）数据大小：以1024字节作为基准（m1）

2048（m2），4096（m4），8192（m8），

512（mf2），256（mf4），128（mf8）

（4）支持mask操作

示例：

代码块

```
tint32m1_t tld_linear_global_m1(const int32_t * op0, uint32_t op1, long op2);
tint32m2_t tld_linear_share_m2(const int32_t * op0, uint32_t op1, long op2);
tint32m4_t tld_blk_global_m4(const int32_t * op0, long op1);
tint32m1_t tld_linear_global_m1_tm(const int32_t * op0, uint32_t op1, long
op2);
```

1.2 store

与load指令相对应，负责数据存储操作，示例如下：

代码块

```
void tst_linear_global_m1(tint32m1_t op0, int32_t * op1, uint32_t op2, long
op3);
void tst_linear_share_m2(tint32m2_t op0, int32_t * op1, uint32_t op2, long
op3);
void tst_blk_global_m4(tint32m4_t op0, int32_t * op1, long op2);
```

```
4   void tst_linear_global_m1_tm(tint32m1_t op0, int32_t * op1, uint32_t op2, long
    op3);
```

1.3 mma

矩阵乘法指令，完成形如：D = A x transpose（B） + D的运算（矩阵乘仅支持K连续排布的模式），一般有：

A shape：m * k

B shape：n * k

D shape：m * n

主要支持功能有：

（1）输出数据类型为f32或s32

（2）输入数据类型为tf32，f16，bf16，fp8，u8，s8，u4，s4

（3）支持数据复用

（4）支持累加、非累加模式

示例如下：

代码块

```
1   tfloat32m2_t tmma(taccfloat32m2_t op0, tbfloat16m1_t op1, tbfloat16m2_t op2);
2   tfloat32m2_t tmma_noacc(tbfloat16m1_t op0, tbfloat16m2_t op1);
3   taccfloat32m2_t tmma_noacc_reused(tbfloat16m1_t op0, tbfloat16m2_t op1);
```

1.4 mva

向量、矩阵乘法指令，与mma指令类似，区别在于：

A shape：1 * k

B shape：n * k

D shape：1 * n

即A为vector，B为matrix，示例如下：

代码块

```
1   tfloat32mf8_t tmva(taccfloat32mf8_t op0, tfloat16mf8_t op1, tfloat16m4_t op2);
2   tfloat32mf8_t tmva_noacc(tfloat16mf8_t op0, tfloat16m4_t op1);
```

```
3   taccfloat32mf8_t tmva_noacc_reused(tfloat16mf8_t op0, tfloat16m4_t op1);
```

1.5 convert

ALU单元指令，完成数据类型转换功能，支持matrix tile和vector tile布局的类型转换，支持特性有：

（1）目标数据类型支持f8_e4m3，f32，bf16，f16，u8，s8，u4，s4等

（2）源数据类型支持f32，bf16，f16，s32等

示例如下：

代码块

```
1   tbfloat16m1_t tcvt_tile_bf16(tfloat32m2_t op0, uint32_t op1);
2   tbfloat16mf8_t tcvt_vec_bf16(tfloat32mf4_t op0, uint32_t op1);
```

1.6 add

ALU单元指令，完成数据累加功能，支持特性有：

（1）Tile reg与Tile reg相加

（2）Tile reg与scalar reg相加

（3）Tile reg与immediate number相加

（4）支持对2个操作数取负数操作（可实现减法）

示例如下：

代码块

```
1   tfloat32m1_t tadd(tfloat32m1_t op0, uint32_t op1);
2   tfloat32m1_t tadd(tfloat32m1_t op0, float op1);
3   tfloat32m1_t tadd(tfloat32m1_t op0, tfloat32m1_t op1);
4   tfloat32m1_t tadd_neg2(tfloat32m1_t op0, tfloat32m1_t op1);
```

1.7 mul

ALU单元指令，完成乘法功能，支持特性与add指令类似，示例如下：

```
tfloat32m1_t tmul(tfloat32m1_t op0, uint32_t op1);
tfloat32m1_t tmul(tfloat32m1_t op0, float op1);
tfloat32mf4_t tmul(tfloat32mf4_t op0, float op1);
tfloat32m1_t tmul(tfloat32m1_t op0, tfloat32m1_t op1);
```

1.8 max/min

ALU单元指令，完成比较功能，支持特性与add指令类似，示例如下：

代码块

```
tfloat32m1_t tmax(tfloat32m1_t op0, tfloat32m1_t op1);
tfloat32m1_t tmin(tfloat32m1_t op0, tfloat32m1_t op1);
```

1.9 fma

ALU单元指令，完成乘加功能（3操作数），支持特性与add指令类似，示例如下：

代码块

```
tbfloat16m1_t tfma_neg1(tbfloat16m1_t op0, tbfloat16m1_t op1, tbfloat16m1_t op2);
tfloat32m1_t tfma(tfloat32m1_t op0, tfloat32m1_t op1, tfloat32m1_t op2);
```

1.10 move

数据搬运指令，支持特性有：

（1）数据由Tile reg搬运至RV Vector Reg

（2）数据由RV Vector Reg搬运至Tile reg

（3）数据广播，由salar广播为Tile reg

代码块

```
vfloat32m1_t tmv_v_f32(tfloat32m1_t op0, unsigned long op1);
tfloat32m1_t tmv_t_f32(tfloat32m1_t op0, uint32_t op1, unsigned long op2);
tfloat32m1_t tmv_t_f32(tfloat32m1_t op0, vfloat32m1_t op1, unsigned long op2);
tfloat32m1_t tmv_t_f32(tfloat32m1_t op0, float op1, unsigned long op2);
```

```
5   tfloat32m1_t tmv_t_f32(uint32_t op0);
```

1.11 wait

Tile同步指令，用于进行多线程间同步，支持特性有：

（1）与acp指令搭配，完成异步拷贝任务

（2）支持global memory与shared memory

示例如下：

代码块

```
1   void twait_load_share(uint32_t op0);
2   void twait_load_global(uint32_t op0);
3   void twait_tacp_cg(uint32_t op0);
```

1.12 async copy

Tile异步拷贝指令，示例如下：

代码块

```
1   void tacp_commit_group();
2   void tacp_tile_srctm(vint32m1_t op0, vint32m1_t op1, const uint8_t * op2);
3   void tacp_convert_srctm(vint32m1_t op0, vint32m1_t op1, const uint8_t * op2);
```

2 常用数据类型

代码块

```
1   tfloat32m1_t，1024字节
2   tfloat32mf2_t，512字节
3
4   tbfloat16m1_t，1024字节
5   tbfloat16mf2_t，512字节
6
7   tfloat16m1_t，1024字节
8   tfloat16mf2_t，512字节
```

```
9
10  tint32m1_t, 1024字节
11  tint32mf2_t, 512字节
12
13  tint8m1_t, 1024字节
14  tint8mf2_t, 512字节
```