A note on the algebra of CuTe Layouts

Jay Shah[†]

1 INTRODUCTION

The core abstraction of NVIDIA's CUTLASS library for high-performance linear algebra is a specific notion of *layout*, introduced as part of its new backend core library CuTe in version 3.0 [1]. CuTe Layouts comprise a convenient formalism for describing and manipulating data of a multi-dimensional nature, such as the values of a matrix or tensor. The goal of this technical note is to study CuTe Layouts from a rigorous, mathematical point of view. Currently, the focus is on articulating sufficient conditions for the operations of complementation and composition to be well-defined, and also to provide explicit closed formulae for them. These operations play an important role in and of themselves, but also jointly in defining the operation of *logical division*. This operation, as well as its relatives such as *zipped division*, plays a critical role in various tiling and slicing operations for CuTe Layouts and Tensors (which are essentially Layouts together with pointers into memory).

This note should be read as complementary to the discussion of layout operations in the CuTe documentation [2]. However, we think that certain portions of that documentation are mathematically vague or false if interpreted literally, which spurred the writing of this note. Most significantly, no discussion of necessary conditions for the operation of composition to be well-defined is given there. This becomes problematic when, for example, it is claimed that composition is left-distributive with respect to concatenation.[1] In code, this is given as a *definition* of composition in the general case. But consider the simple example of

```
Layout A = make_layout(make_shape(_6{},_2{}),make_stride(_1{},_7{}));
Layout B = make_layout(make_shape(_3{},_2{}),make_stride(_2{},_3{}));
Layout C = composition(A,B);
```

Then when running with CUTLASS 3.3, the layout C evaluates to

```
(_3,_2):(_2,_3)
```

since C is defined according to the supposed left-distributivity property. But note that C doesn't actually describe the composition of A and B in terms of the associated layout functions f_A, f_B, and f_C. Indeed, we have that

$$f_C(5) = f_C(2) + f_C(3) = 4 + 3 = 7,$$

whereas

$$(f_A \circ f_B)(5) = f_A(7) = f_A(1) + f_A(6) = 1 + 7 = 8.$$

Actually, in this case $A \circ B$ will not be well-defined as a layout, even though for the separate modes B_0 and B_1 of B, the compositions $A \circ B_0$ and $A \circ B_1$ are well-defined. This "overflow" issue occurs since a certain disjointness condition is violated, which we articulate as Definition 2.17. Of course, in practice the programmer would not consider such a composition to begin with, but we hope that our note can serve as an all-purpose reference for when such operations are meant to be valid. However, we emphasize that the treatment of layouts given in this note is entirely implementation-agnostic.

The contents of the current note form a self-contained body of work as it stands, although it will appear unmotivated if the reader doesn't already have prior experience working with CuTe Layouts. We anticipate adding to this document as the need arises, or if elaborations of other aspects of layout algebra are desired by CUTLASS/CuTe developers.

[1] Item (3) in "Rules for computing composition" from [2].

[†] Colfax Research. A copy of this note is available at https://research.colfax-intl.com/.
Date: January 16, 2024. *Email*: jayhshah@colfax-intl.com.

2 LAYOUT ALGEBRA

Definition 2.1. A *layout* L is a pair of positive[2] integer tuples $\mathbf{S}$ and $\mathbf{D}$ of matching dimensions. We call $\mathbf{S}$ the *shape* and $\mathbf{D}$ the *stride*. We write $L = \mathbf{S} : \mathbf{D}$.

From now on in this note, we assume that layouts are flattened (i.e., internal parentheses for S and D have been removed); this won't change the semantics of the operations that we consider. Let's first introduce some basic terminology:

Definition 2.2. Let $\alpha \geq 0$ be an integer and $L = \mathbf{S} : \mathbf{D} = (M_0, ..., M_\alpha) : (d_0, ..., d_\alpha)$ be a layout. Then:

- The *size* of L is the product $M = M_0 \cdot ... \cdot M_\alpha$.
- The *length* of L is the integer $\alpha + 1$.
- A *mode* of L is one of the entries $(M_k) : (d_k)$ for $0 \leq k \leq \alpha$. We may regard this as a length 1 layout.

Given two layouts $L = \mathbf{S} : \mathbf{D}$ and $L' = \mathbf{S}' : \mathbf{D}'$, let $\mathbf{S}''$ and $\mathbf{D}''$ be the shape and stride tuples given by (the flattening of) $(\mathbf{S}, \mathbf{S}')$ and $(\mathbf{D}, \mathbf{D}')$. Then the *concatenation* of L and L' is given by the layout

$$(L, L') \coloneqq \mathbf{S}'' : \mathbf{D}'',$$

and we say that (L, L') is decomposed by L and L'. Inductively, given layouts $L_1, ..., L_N$, we can then form the concatenated layout $(L_1, ..., L_N)$. Conversely, given L a layout, L is maximally decomposed by its modes.

To each layout L, we can associate a function as follows. Let $\mathbf{S} = (M_0, ..., M_\alpha)$ and $\mathbf{D} = (d_0, ..., d_\alpha)$ be the respective shape and stride tuples for L. Let $M = M_0 \cdot M_1 \cdot ... \cdot M_\alpha$ be the size of L and let $[0, M) \subset \mathbb{N}$ be the subset of the natural numbers given by $\{0, ..., M-1\}$. Then we have an isomorphism

$$\iota : [0, M) \cong [0, M_0) \times [0, M_1) \times \ldots \times [0, M_\alpha)$$

given by $x \mapsto (x \bmod M_0, \left\lfloor \frac{x}{M_0} \right\rfloor \bmod M_1, \ \ldots, \left\lfloor \frac{x}{M_0 \cdot ... \cdot M_{\alpha-1}} \right\rfloor \bmod M_\alpha)$.

Definition 2.3. Given a layout L, its *layout function* $f_L : [0, M) \to \mathbb{N}$ is defined to be the composite

$$[0, M) \cong [0, M_0) \times \ldots \times [0, M_\alpha) \subset \mathbb{N}^{\times(\alpha+1)} \xrightarrow{(\cdot d_0, ..., \cdot d_\alpha)} \mathbb{N}^{\times(\alpha+1)} \xrightarrow{+} \mathbb{N}.$$

In other words, f_L is the composition of the multi-linear function

$$[0, M_0) \times \ldots \times [0, M_\alpha) \to \mathbb{N}, \quad (x_0, ..., x_\alpha) \mapsto d_0 x_0 + ... + d_\alpha x_\alpha,$$

determined by the stride, with the isomorphism ι, determined by the shape.

We then let $\widehat{f_L} : \mathbb{N} \to \mathbb{N}$ be the extension of f_L defined by replacing M_α by ∞, i.e., the composite

$$\mathbb{N} \cong [0, M_0) \times \ldots \times [0, M_{\alpha-1}) \times \mathbb{N} \subset \mathbb{N}^{\times(\alpha+1)} \xrightarrow{(\cdot d_0, ..., \cdot d_\alpha)} \mathbb{N}^{\times(\alpha+1)} \xrightarrow{+} \mathbb{N}$$

where the first isomorphism is the extension $\widehat{\iota}$ of ι given by

$$x \mapsto (x \bmod M_0, \left\lfloor \frac{x}{M_0} \right\rfloor \bmod M_1, \ \ldots, \left\lfloor \frac{x}{M_0 \cdot ... \cdot M_{\alpha-2}} \right\rfloor \bmod M_{\alpha-1}, \left\lfloor \frac{x}{M_0 \cdot ... \cdot M_{\alpha-1}} \right\rfloor).$$

[2]For our purposes, we ignore the empty layout as well as zero strides.

2.1 Complementation

In this subsection, we define the notion of the *complement* of a layout A with respect to a given integer M, under certain assumptions.

Definition 2.4. Let $A = (N_0, ..., N_\alpha) : (d_0, ..., d_\alpha)$ be a layout. We say that A is *sorted* if $d_0 \leq ... \leq d_\alpha$ and for every $i < j$ such that $d_i = d_j$, $N_i \leq N_j$.

Definition 2.5. Let $A = (N_0, ..., N_\alpha) : (d_0, ..., d_\alpha)$ be a layout and M a positive integer. Suppose without loss of generality that A is sorted; if not, replace A with a permutation of itself that is sorted. Then we say that the pair $\{A, M\}$ is *admissible for complementation* (or simply admissible) if:

- For all $1 \leq i \leq \alpha$, the product $N_{i-1}d_{i-1}$ divides d_i.
- The product $N_\alpha d_\alpha$ divides M.

Definition 2.6. Let $A = (N_0, ..., N_\alpha) : (d_0, ..., d_\alpha)$ be a layout and M a positive integer. Suppose that $\{A, M\}$ is admissible for complementation and reindex A so that it is sorted. Then the complement of $\{A, M\}$ is defined to be the layout

$$\text{complement}(A, M) = (d_0, \frac{d_1}{N_0 d_0}, \frac{d_2}{N_1 d_1}, \ldots, \frac{M}{N_\alpha d_\alpha}) : (1, N_0 d_0, N_1 d_1, \ldots, N_\alpha d_\alpha).$$

Note that by definition, the complement of A (taken with respect to some integer M) is insensitive to permutations of A. Moreover, its layout function is strictly increasing.

The following proposition explains the sense in which Definition 2.6 is taking a complement.

Proposition 2.7. *Let $\{A = (N_0, ..., N_\alpha) : (d_0, ..., d_\alpha), M\}$ be an admissible pair and $B = \text{complement}(A, M)$. Let $C = (A, B)$ be the concatenated layout. Then the size of C is M and $f_C : [0, M) \to \mathbb{N}$ restricts to a bijection $[0, M) \cong [0, M)$.*

Proof. Since $\text{size}(A) \cdot \text{size}(B) = M$, we see that the domain of f_C is indeed $[0, M)$. Note that the image of f_C is the same as that of $f_{C'}$ for any permutation C' of C. Therefore, when computing the image of f_C we may sort C so that the strides are in non-decreasing order, as well as reindex A so that it is sorted. So after reindexing A, let

$$C' = (d_0, N_0, \frac{d_1}{N_0 d_0}, N_1, \frac{d_2}{N_1 d_1} ..., N_\alpha, \frac{M}{N_\alpha d_\alpha}) : (1, d_0, N_0 d_0, d_1, N_1 d_1, ..., d_\alpha, N_\alpha d_\alpha).$$

Then we may write

$$C' = (r_0, r_1, r_2, ..., r_\beta) : (1, r_0, r_0 r_1, ..., r_0 ... r_{\beta-1})$$

for $\beta = 2\alpha + 1$, and the maximum value that f_C attains is computed as

$$(r_0 - 1) + r_0(r_1 - 1) + (r_0 r_1)(r_2 - 1) + ... + (r_0 ... r_{\beta-1})(r_\beta - 1) = r_0 r_1 ... r_\beta - 1 = M - 1.$$

To establish the bijectivity assertion, it then suffices to show that $f_{C'}$ is injective. For this, suppose that $x, y \in [0, M)$ so that $f_{C'}(x) = f_{C'}(y)$, and let $(x_0, ..., x_\beta)$ and $(y_0, ..., y_\beta)$ be their coordinate vectors with respect to C'. Expanding the terms in the equality we get

$$x_0 + r_0 x_1 + (r_0 r_1) x_2 + ... + (r_0 ... r_{\beta-1}) x_\beta = y_0 + r_0 y_1 + (r_0 r_1) y_2 + ... + (r_0 ... r_{\beta-1}) y_\beta.$$

We show by induction that $x_i = y_i$ for all $i \in \{0, ..., \beta\}$, which will complete the proof. Firstly, taking both sides mod r_0 shows that $x_0 = y_0$ since both lie in $[0, r_0)$. Now suppose by induction that given $0 < i \leq \beta$, for all $j < i$ we have $x_j = y_j$. Then we can reduce the expression to

$$(r_0 ... r_{i-1}) x_i + ... + (r_0 ... r_{\beta-1}) x_\beta = (r_0 ... r_{i-1}) y_i + ... + (r_0 ... r_{\beta-1}) y_\beta.$$

Taking this equation mod $r_0 ... r_i$ and dividing by $(r_0 ... r_{i-1})$ shows that $x_i = y_i$, since we know both lie in $[0, r_i)$. □

Corollary 2.8. *In the setting of Proposition 2.7, let* $I = [0, \text{size}(A)) = [0, N_0 ... N_\alpha)$ *be the domain of* f_A*. Then*

$$f_A(I) \cap \widehat{f_B}(I) = \{0\}.$$

In other words, $\widehat{f_A}$ *and* $\widehat{f_B}$ *have disjoint image when restricted to the domain of* f_A*, apart from* 0*.*

Proof. Let $J = [0, \text{size}(B)) = [0, M/(N_0 ... N_\alpha))$. By Proposition 2.7, we have that

$$f_A(I \cap J) \cap f_B(I \cap J) = \{0\}.$$

It remains to consider values of the extended function $\widehat{f_B}$ on integers that might lie in I but not J. But $\widehat{f_B}$ is a strictly increasing function, $\widehat{f_B}(\text{size}(B)) = M$, and the largest value attained by f_A satisfies the inequality

$$\begin{aligned}(N_0 - 1)d_0 + (N_1 - 1)d_1 + ... + (N_\alpha - 1)d_\alpha &< d_1 + (N_1 - 1)d_1 + (N_2 - 1)d_2 + ... + (N_\alpha - 1)d_\alpha \\ &\leq d_2 + (N_2 - 1)d_2 + ... + (N_\alpha - 1)d_\alpha \leq ... \\ &\leq d_\alpha + (N_\alpha - 1)d_\alpha \leq M.\end{aligned}$$

□

Remark 2.9. The CuTe documentation [2] stipulates that the complement B of a layout A with respect to an integer M should satisfy three properties:

(1) A and B are *disjoint* in the sense that $f_A(x) \neq f_B(x)$ for all $x \neq 0$ in the domain of f_A;

(2) B is *ordered* in the sense that f_B is a strictly increasing function;

(3) B is *bounded* by M in the sense that $\text{size}(B) \geq M/\text{size}(A)$ and $\text{cosize}(B) \leq \left\lfloor \frac{M}{\text{cosize}(A)} \right\rfloor \cdot \text{cosize}(A)$. Here, we let the cosize of a layout A be given by $f_A(\text{size}(A) - 1) + 1$.

We observe that all these properties are satisfied by the definition of complement given in Definition 2.6 for $\{A, M\}$ admissible. (1) follows from Corollary 2.8.[3] (2) follows by definition of the complement as we noted above. Finally, for (3) we have that $\text{size}(B) = M/\text{size}(A)$ and

$$\begin{aligned}\text{cosize}(B) &= 1 + (d_0 - 1) + \left(\frac{d_1}{N_0 d_0} - 1\right) N_0 d_0 + ... + \left(\frac{M}{N_\alpha d_\alpha} - 1\right) N_\alpha d_\alpha \\ &= d_0 + (d_1 - N_0 d_0) + ... + (d_\alpha - N_{\alpha-1} d_{\alpha-1}) + M - N_\alpha d_\alpha \\ &= M - ((N_0 - 1)d_0 + ... + (N_\alpha - 1)d_\alpha) \\ &= M - (\text{cosize}(A) - 1),\end{aligned}$$

where we reindexed A according to its sort for the intermediate terms; this doesn't change the final equality. Therefore, the inequality to check for the cosizes becomes

$$\frac{M}{\text{cosize}(A)} - 1 + \frac{1}{\text{cosize}(A)} \leq \left\lfloor \frac{M}{\text{cosize}(A)} \right\rfloor,$$

which holds for any pair of positive integers.

Example 2.10. We give two examples in CUTLASS 3.3 for when CuTe's complement method can be evaluated but has potentially undesired behavior. Consider the layout $A = (4) : (2)$ and $M = 19$, so we don't have that $\{A, M\}$ is admissible. Then `complement(A,M)` evaluates to

```
(_2,_3):(_1,_8)
```

[3] Corollary 2.8 is actually stronger since it concerns disjointness of the images.

However, in this case $\mathrm{cosize}(B) = 18$, whereas $\mathrm{cosize}(A) = 7$ and thus

$$\left\lfloor \frac{M}{\mathrm{cosize}(A)} \right\rfloor \cdot \mathrm{cosize}(A) = \left\lfloor \frac{19}{7} \right\rfloor \cdot 7 = 2 \cdot 7 = 14.$$

Now consider $A = (2,2) : (2,3)$ and $M = 19$. Then `complement(A,M)` evaluates to

```
(_2,_0,_4):(_1,_4,_6)
```

which is the empty layout (with $\mathrm{size}(B) = 0$), since 0 occurs in its shape tuple.

2.2 Composition

We next discuss the operation of composition of layouts A and B. **For simplicity, we suppose that the shape tuples contain no integers equal to** 1; stripping out these modes doesn't change the associated layout function. The goal here is to produce a layout, denoted $A \circ B$, whose associated function $f_{A\circ B}$ identifies with the composition $\widehat{f_A} \circ f_B$. In general, we need conditions in order to be able to define $A \circ B$.

Definition 2.11. Let $M, d > 0$ be positive integers and let $M = M_0 \cdot M_1 \cdot \ldots \cdot M_\alpha$ be a given factorization of M by integers $M_k > 1$. Replacing M_α by ∞, let

$$\widehat{M} = M_0 \cdot M_1 \cdot \ldots \cdot M_{\alpha-1} \cdot \infty$$

and consider ∞ to be divisible by every positive integer. We say that M is *left divisible* by d (implicitly, with respect to the given factorization) if there exists $0 \leq i \leq \alpha$ such that:

(1) $M_0 \ldots M_{i-1}$ divides d.[4]

(2) Supposing (1), let $c = d/(M_0 \ldots M_{i-1})$.[5] Then if $i < \alpha$, we require in addition that $1 \leq c < M_i$.

(3) For (2) in the case $i < \alpha$, we require in addition that c also divides M_i.

Note that i is necessarily unique if it exists. In this case, we will refer to i as the *division index* and write $\widehat{M} = d \cdot \widehat{M'}$. Moreover, we will endow $\widehat{M'}$ with the following induced factorization:

(a) If $0 \leq i < \alpha$, then $\widehat{M'} = M'_0 \cdot \ldots \cdot M'_{\alpha-i-1} \cdot \infty$ with $M'_0 = M_i/c > 1$ and $M'_j = M_{i+j}$ for $0 < j < \alpha - i$.

(b) If $i = \alpha$, then $\widehat{M} = d \cdot \infty$ and we will let $\widehat{M'} = \infty$.

Furthermore, we say that M is *weakly left divisible* by d if there exists $0 \leq i \leq \alpha$ such that the above conditions (1) and (2) hold, but not necessarily (3). Then we still call the (necessarily unique) i the division index as before, but we no longer have the factorization of $\widehat{M}$.

Note that in Definition 2.11, the term $\widehat{M'}$ with its induced factorization can itself be considered for left divisibility or weak left divisibility (with the step of replacing the last factor by ∞ now being superfluous).

We first consider composition in the restricted case of length 1 layouts for the second layout. To this end, we have the following notion of "admissibility for composition":

Definition 2.12. Let $\mathbf{S} = (M_0, \ldots, M_\alpha)$ be a shape tuple, let $M = M_0 \ldots M_\alpha$, and let $B = (N) : (r)$ be a layout of length 1. Then we say that the pair $\{\mathbf{S}, B\}$ is *admissible for composition* (or simply admissible) if:

(1) M is left divisible by r. Write $\widehat{M} = r \cdot \widehat{M'}$.

(2) With respect to its induced factorization, $\widehat{M'}$ is weakly left divisible by N.

[4] If $i = 0$, we regard the empty product as equal to 1, so that this is no condition.
[5] If $i = 0$, this means that $c = d$.

The idea of admissibility is that the composition $A \circ B$ of layouts will entail "dividing B along the modes of A". More precisely, we have the following:

Definition 2.13. Suppose that $\mathbf{S} = (M_0, ..., M_\alpha)$ is a shape tuple and $B = (N) : (r)$ is a layout of length 1 such that $\{\mathbf{S}, B\}$ is admissible. Let $\mathbf{D} = (d_0, ..., d_\alpha)$ be any stride tuple and let $A = \mathbf{S} : \mathbf{D}$.

As in Definition 2.11, let $M = M_0 \cdot ... \cdot M_\alpha$ and $\widehat{M} = r \cdot \widehat{M'}$ with division index $0 \leq i \leq \alpha$. We separate the definition of $A \circ B$ into two cases. First suppose that $0 \leq i < \alpha$, so that

$$r = M_0 \cdot ... \cdot M_{i-1} \cdot c, \quad \widehat{M'} = M_i/c \cdot ... \cdot \infty.$$

Then if $N \leq M_i/c$, we let $A \circ B = (N) : (cd_i)$. Otherwise, we have that $N = M_i/c \cdot ... \cdot M_{j-1} \cdot c'$ (where $c' < M_j$ if $j \neq \alpha$), and we let

$$A \circ B = \begin{cases} (M_i/c, M_{i+1}, ..., M_{j-1}, c') : (cd_i, d_{i+1}, ..., d_{j-1}, d_j) & \text{if } c' > 1; \\ (M_i/c, M_{i+1}, ..., M_{j-1}) : (cd_i, d_{i+1}, ..., d_{j-1}) & \text{if } c' = 1. \end{cases}$$

If instead $i = \alpha$, then we have $r = M_0 \cdot ... \cdot M_{\alpha-1} \cdot c$ as before but $\widehat{M'} = \infty$, and we let $A \circ B = (N) : (cd_\alpha)$.

Note that by definition the size of $A \circ B$ always equals that of B. We then have the following soundness proposition for Definition 2.13. In the proof, we will use the following notation: for a given index $0 \leq k \leq \alpha$, let $\delta_k \in \mathbb{N}^{\times(\alpha+1)}$ denote the coordinate that is zero everywhere except in the kth position, where it is 1.

Proposition 2.14. *In the situation of Definition 2.13, we have that $f_{A \circ B} = \widehat{f_A} \circ f_B$.*

Proof. We carry over notation from Definition 2.13. Then with respect to the isomorphism

$$\widehat{\imath} : \mathbb{N} \cong [0, M_0) \times ... \times [0, M_{\alpha-1}) \times \mathbb{N}$$

of Definition 2.3, we have that r is sent to $c \cdot \delta_i$. Thus, we see that

$$(\widehat{f_A} \circ f_B)(1) = cd_i = f_{A \circ B}(1).$$

In the cases of $i < \alpha$ and $N \leq M_i/c$ or $i = \alpha$, this already suffices to show $f_{A \circ B} = \widehat{f_A} \circ f_B$. In the remaining case $i < \alpha$ and $N = M_i/c \cdot ... \cdot M_{j-1} \cdot c'$, note that

$$\widehat{\imath}((M_i/c)r) = \delta_{i+1}, \ \widehat{\imath}(M_{i+1}(M_i/c)r) = \delta_{i+2}, \ ..., \ \widehat{\imath}\big(M_{j-1}...M_{i+1}(M_i/c)r\big) = \delta_j.$$

Therefore, we see that $f_{A \circ B}$ and $\widehat{f_A} \circ f_B$ agree on values $\{1, \ M_i/c, \ M_{i+1}(M_i/c), \ ..., \ M_{j-1}...M_{i+1}(M_i/c)\}$ (or drop the last term if $c' = 1$). In view of the multi-linearity properties of both functions,[6] this implies that $f_{A \circ B} = \widehat{f_A} \circ f_B$. □

Example 2.15. Let $A = (M_0, ..., M_\alpha) : (d_0, ..., d_\alpha)$ be any layout. For $i = 0$, let $B_0 = (M_0) : (1)$, and for $0 < i \leq \alpha$, let $B_i = (M_i) : (M_0 \cdot ... \cdot M_{i-1})$. Then $A \circ B_i = (M_i) : (d_i)$.

To extend from the case of length 1 layouts to general layouts for the term B in a putative composition $A \circ B$, we will write $B = (B_0, ..., B_\beta)$ as a concatenation of its modes and then concatenate the resulting compositions $A \circ B_0, ..., A \circ B_\beta$. For this to yield a correct result in general, we need to avoid potential collisions.

Definition 2.16. In the situation of Definition 2.12, let $f_B : [0, N) \to \mathbb{N}$ be the layout function, and let $I = [r, r(N-1)]$ be the interval given by the convex closure of the image $f_B([1, N))$. Let $M' = M_0...M_{\alpha-1}$ and $J = I \cap [1, M')$ (so $J = \emptyset$ if $\alpha = 0$). Then the *interval of definition* for $\{\mathbf{S}, B\}$ is J.

Definition 2.17. Let $\mathbf{S} = (M_0, ..., M_\alpha)$ be a shape tuple, let $B = (N_0, ..., N_\beta) : (r_0, ..., r_\beta)$ be a layout, and let $B_k = (N_k) : (r_k)$ for $0 \leq k \leq \beta$. Then we say that the pair $\{\mathbf{S}, B\}$ is *admissible for composition* if:

[6] The reader should compare this situation with the obstacle that arises in the proof of Theorem 2.18 below.

(1) For all $0 \leq k \leq \beta$, the pair $\{\mathbf{S}, B_k\}$ is admissible for composition in the sense of Definition 2.12.

(2) The intervals of definition for the pairs $\{\mathbf{S}, B_k\}_{0 \leq k \leq \beta}$ are disjoint.

In this case, if $\mathbf{D} = (d_0, ..., d_\alpha)$ is any stride tuple and $A = \mathbf{S} : \mathbf{D}$, then we define the *composition* $A \circ B$ to be the concatenated layout

$$A \circ B \coloneqq (A \circ B_0, A \circ B_1, ..., A \circ B_\beta)$$

where each $A \circ B_k$ is defined as in Definition 2.13.

We have the following soundness theorem to validate Definition 2.17.

Theorem 2.18. *In the situation of Definition 2.17, we have that* $f_{A \circ B} = \widehat{f_A} \circ f_B$.

Proof. By Proposition 2.14, we have that for all $0 \leq k \leq \beta$, the equality $f_{A \circ B_k} = \widehat{f_A} \circ f_{B_k}$ of functions holds on the domain $[0, \text{size}(B_k))$. By Lemma 2.19, we have that the following diagram commutes:

$$\begin{array}{ccc} [0, \text{size}(B)) & \xrightarrow[\cong]{\iota} & [0, \text{size}(B_0)) \times ... \times [0, \text{size}(B_\beta)) \\ \downarrow{\scriptstyle f_{A \circ B}} & & \downarrow{\scriptstyle (f_{A \circ B_0}, ..., f_{A \circ B_\beta})} \\ \mathbb{N} & \xleftarrow{+} & \mathbb{N} \times ... \times \mathbb{N}. \end{array}$$

It then suffices to see that the analogous diagram with $\widehat{f_A} \circ f_B$ commutes, i.e. for the diagram

$$\begin{array}{ccc} [0, \text{size}(B)) & \xrightarrow[\cong]{\iota} & [0, \text{size}(B_0)) \times ... \times [0, \text{size}(B_\beta)) \\ \downarrow{\scriptstyle \widehat{f_A} \circ f_B} & & \downarrow{\scriptstyle (\widehat{f_A} \circ f_{B_0}, ..., \widehat{f_A} \circ f_{B_\beta})} \\ \mathbb{N} & \xleftarrow{+} & \mathbb{N} \times ... \times \mathbb{N}. \end{array}$$

Breaking out the composition, we may factor this diagram as

$$\begin{array}{ccc} [0, \text{size}(B)) & \xrightarrow[\cong]{\iota} & [0, \text{size}(B_0)) \times ... \times [0, \text{size}(B_\beta)) \\ \downarrow{\scriptstyle f_B} & & \downarrow{\scriptstyle (f_{B_0}, ..., f_{B_\beta})} \\ \mathbb{N} & \xleftarrow{+} & \mathbb{N} \times ... \times \mathbb{N} \\ \downarrow{\scriptstyle \widehat{f_A}} & & \downarrow{\scriptstyle (\widehat{f_A}, ..., \widehat{f_A})} \\ \mathbb{N} & \xleftarrow{+} & \mathbb{N} \times ... \times \mathbb{N} \end{array}$$

where the upper square commutes, again by Lemma 2.19. **Note that the bottom square does not commute in general** (i.e., the function $\widehat{f_A} : \mathbb{N} \to \mathbb{N}$ itself is not generally additive). However, with respect to the factorization

$$\widehat{f_A} : \mathbb{N} \xrightarrow{\cong} [0, M_0) \times ... \times [0, M_{\alpha-1}) \times \mathbb{N} \xrightarrow{(d_0, ..., d_\alpha)} \mathbb{N} \times ... \times \mathbb{N} \xrightarrow{+} \mathbb{N},$$

our assumption of disjoint intervals of definition ensures that the images of the maps $f_{B_0}, ..., f_{B_\beta}$ are disjoint when intersected with $[0, M_0) \times ... \times [0, M_{\alpha-1}) - \{0\}$. For additivity, it now suffices to check that there do not exist distinct B_k, B_l and non-zero $x \in \text{im}\left(f_{B_k}\right)$, $y \in \text{im}\left(f_{B_l}\right)$ that have coordinates $x_i, y_i \in [0, M_i)$ for some $0 \leq i < \alpha$ such that $x_i + y_i \geq M_i$; if not, we may have that

$$\widehat{f_A}(x + y) \neq \widehat{f_A}(x) + \widehat{f_A}(y)$$

due to overflow in the ith coordinate, because the strides for the layout A can be arbitrary. Now let w_{i_0} and z_{j_0} be the leftmost non-zero coordinates of $f_{B_k}(1)$ and $f_{B_l}(1)$, respectively. If either of the indices i_0 or j_0 equal α then we are already done. Otherwise, we have that $w_{i_0} \leq M_{i_0}/2$ and $z_{j_0} \leq M_{j_0}/2$ from the left divisibility assumption. Moreover, the coordinates of subsequent values of f_{B_k} and f_{B_l} will increment by multiples of w_{i_0} and z_{j_0} in indices i_0 and j_0, by increments of 1 for indices greater than i_0 and j_0 up to that occupied by the maximum value, and

zero elsewhere. Finally, by disjointness[7] we have that either $f_{B_l}(1)$ is strictly greater than the maximum value attained by f_{B_k} or vice-versa. Putting this all together, we see that disjointness of the intervals of definition rules out the possibility of overflow.

We conclude that when restricted to the image of $(f_{B_0}, ..., f_{B_\beta})$, we do have that $\widehat{f_A}$ distributes over addition, which completes the proof. □

We used the following lemma about concatenated layouts in the proof of Theorem 2.18.

Lemma 2.19. *Let $C = (C_0, ..., C_\gamma)$ be a concatenated layout. Let*

$$\iota : [0, \text{size}(C)) \cong [0, \text{size}(C_0)) \times ... \times [0, \text{size}(C_\gamma))$$

be the usual isomorphism (as in Definition 2.3). Then the following diagram commutes:

$$\begin{array}{ccc} [0, \text{size}(C)) & \xrightarrow[\cong]{\iota} & [0, \text{size}(C_0)) \times ... \times [0, \text{size}(C_\gamma)) \\ {\scriptstyle f_C}\downarrow & & \downarrow{\scriptstyle (f_{C_0}, ..., f_{C_\gamma})} \\ \mathbb{N} & \xleftarrow{\quad + \quad} & \mathbb{N} \times ... \times \mathbb{N} \end{array}$$

Proof. If $C_0, ..., C_\gamma$ are all length 1 layouts, then this is immediate from the definition. In general, we can take the maximal decomposition $C = (C'_0, ..., C'_{\gamma'})$ where all the C'_j are length 1 layouts and $\gamma' + 1$ is the length of C. Then the C_i will be decomposed by disjoint and convex collections of the C'_j in order, and we may place the diagram in question into the larger diagram

$$\begin{array}{ccccc} [0, \text{size}(C)) & \xrightarrow[\cong]{\iota} & [0, \text{size}(C_0)) \times ... \times [0, \text{size}(C_\gamma)) & \xrightarrow[\cong]{(\iota_0, \ ..., \ \iota_\gamma)} & [0, \text{size}(C'_0)) \times ... \times [0, \text{size}(C'_{\gamma'})) \\ {\scriptstyle f_C}\downarrow & & \downarrow{\scriptstyle (f_{C_0}, ..., f_{C_\gamma})} & & \downarrow{\scriptstyle (f_{C'_0}, ..., f_{C'_{\gamma'}})} \\ \mathbb{N} & \xleftarrow{\quad + \quad} & \mathbb{N}^{\times(\gamma+1)} & \xleftarrow{\quad (+, \ ..., \ +) \quad} & \mathbb{N}^{\times(\gamma'+1)}. \end{array}$$

Here, the maps $\iota_0, ..., \iota_\gamma$ are the usual isomorphisms mapping the intervals $[0, \text{size}(C_i))$ to their corresponding decompositions in terms of products of the intervals $[0, \text{size}(C'_j))$. Now observe that the composite map $(\iota_0, ..., \iota_\gamma) \circ \iota$ is also the usual isomorphism with respect to the maximal decomposition of C. Therefore, by definition the outer rectangle and righthand square commute, hence the lefthand square commutes. □

Example 2.20. As in Example 2.15, let $A = \mathbf{S} : \mathbf{D} = (M_0, ..., M_\alpha) : (d_0, ..., d_\alpha)$ be an arbitrary layout and

$$B_0 = (M_0) : (1), \ B_1 = (M_1) : (M_0), \ ..., \ B_\alpha = (M_\alpha) : (M_0 ... M_{\alpha-1}).$$

Let $U \subset [0, \alpha]$ be any nonempty subset. Then for the collection of pairs $\{\mathbf{S}, B_k\}_{k \in U}$, the intervals of definition will be disjoint. Therefore, if we let B_U be the concatenation of the B_k for $k \in U$, then the pair $\{\mathbf{S}, B_U\}$ is admissible for composition. Explicitly, if we write $U = \{i_0, ..., i_\gamma\}$, then we have

$$A \circ B_U = (M_{i_0}, ..., M_{i_\gamma}) : (d_{i_0}, ..., d_{i_\gamma}).$$

We may think of precomposition with B_U as a *projector* to the modes of A with indices in U.

Warning 2.21. The conditions articulated in Definition 2.12 for single-mode admissibility are more relaxed than the static assert checks carried out in CUTLASS itself.[8] Namely, our condition (1) is identical to a condition checked by CUTLASS, whereas for condition (2), our requirement of weak left divisibility is substituted by (ordinary) left divisibility in CUTLASS. For example, consider the layouts $A = (4, 6, 8, 10) : (2, 3, 5, 7)$ and $B = (6) : (12)$. Then attempting to compute the composition $C = A \circ B$ yields the error message "static assertion failed with "Static shape_div failure"" in CUTLASS, whereas according to our rules we would compute C as $(2, 3) : (9, 5)$.

[7] In this part of the proof it is essential that we took the convex closure of the image in Definition 2.16.

[8] We thank Cris Cecka for a helpful conversation on this point.

2.3 Logical Division

With these preliminaries in place, we can define the operation of logical division.

Definition 2.22. Let $A = \mathbf{S} : \mathbf{D}$ and B be layouts, and let M be the size of A. Suppose that the pairs $\{B, M\}$ and $\{\mathbf{S}, B\}$ are admissible (for complementation and composition, respectively). Then we define the *logical division* A/B to be the layout

$$A/B \coloneqq A \circ (B, \text{complement}(B, M)).$$

Implicit in Definition 2.22 is the following lemma:

Lemma 2.23. *Suppose $A = \mathbf{S} : \mathbf{D}$, $M = \text{size}(A)$, and B are as in Definition 2.22. Then $\{\mathbf{S}, (B, \text{complement}(B, M))\}$ is admissible for composition.*

Proof. Write $A = \mathbf{S} : \mathbf{D} = (M_0, ..., M_\alpha) : (d_0, ..., d_\alpha)$ and $B = (N_0, ..., N_\beta) : (r_0, ..., r_\beta)$. Let

$$\varphi : [0, \beta] \xrightarrow{\cong} [0, \beta]$$

be the automorphism such that $B^\varphi \coloneqq (N_{\varphi(0)}, ..., N_{\varphi(\beta)}) : (r_{\varphi(0)}, ..., r_{\varphi(\beta)})$ is sorted. Then by definition,

$$\text{complement}(B, M) = \left(r_{\varphi(0)}, \frac{r_{\varphi(1)}}{N_{\varphi(0)} r_{\varphi(0)}}, ..., \frac{M}{N_{\varphi(\beta)} r_{\varphi(\beta)}} \right) : \left(1, N_{\varphi(0)} r_{\varphi(0)}, ..., N_{\varphi(\beta)} r_{\varphi(\beta)}\right).$$

Now write

$$B'_0 = \left(r_{\varphi(0)}\right) : (1)\,,\ B'_1 = \left(\frac{r_{\varphi(1)}}{N_{\varphi(0)} r_{\varphi(0)}} \right) : \left(N_{\varphi(0)} r_{\varphi(0)}\right),\ \ldots\ ,\ B'_\beta = \left(\frac{M}{N_{\varphi(\beta)} r_{\varphi(\beta)}} \right) : \left(N_{\varphi(\beta)} r_{\varphi(\beta)}\right)$$

for the length 1 layouts that comprise complement(B, M). We first claim that the pairs $\{\mathbf{S}, B'_k\}$ for $0 \le k \le \beta$ are all admissible for composition. By assumption, we have that M is left divisible by $r_{\varphi(k)}$ and its remainder is then weakly left divisible by $N_{\varphi(k)}$, for all $0 \le k \le \beta$. But since $r_{\varphi(k)} N_{\varphi(k)} | r_{\varphi(k+1)}$ for all $0 \le k < \beta$ and $M = \text{size}(A)$, the additional divisibility condition (3) in Definition 2.11 needed to promote weak left divisibility to left divisibility is necessarily satisfied for all the $N_{\varphi(k)}$ terms. Therefore, we deduce that the pairs $\{\mathbf{S}, B'_k\}$ are indeed all admissible. Now by Proposition 2.7, we see that the additional disjointness assumption is satisfied so that $\{\mathbf{S}, (B, \text{complement}(B, M))\}$ is admissible for composition. □

This concludes our current treatment of logical division. For the time being, we leave further discussion of examples of logical division to the CuTe documentation.

3 PERMUTATIONS EXPRESSIBLE AS LAYOUT FUNCTIONS

In this section, we explain how to retrieve all permutations that are expressible as layout functions in a structured way (for some more precise motivation, we refer to Remark 3.16 below). We will assume that the reader is familiar with the basic language of category theory, which is convenient for describing the algebraic structure of "ordered factorizations" that naturally appears here.

Definition 3.1. We define the set ob(**Fact**) of ordered factorizations to consist of all expressions $[p_1 ... p_k]$ where $k \ge 0$ and the p_i are primes (not necessarily distinct). The case $k = 0$ corresponds to the empty factorization, which we denote as $[]$.

Example 3.2. The set ob(**Fact**) includes expressions such as [], [2], [3], [22], [23], [32], [232], etc.

Notation 3.3. Let $\underline{k}$ denote the set $\{1, 2, ..., k\}$ consisting of k elements. (If $k = 0$, then $\underline{0} = \emptyset$ is the empty set.)

Definition 3.4. We define the category **Fact** of ordered factorizations as follows:

(1) ob(**Fact**) is the set of objects of **Fact**.

(2) For every expression $E = [p_1 p_2 ... p_k]$ in ob(**Fact**) and every morphism of finite sets $\alpha : \underline{n} \to \underline{k}$, we have a morphism
$$E^\alpha = [p_{\alpha(1)} p_{\alpha(2)} ... p_{\alpha(n)}] \xrightarrow{\alpha_E} E = [p_1 p_2 ... p_k]$$
in **Fact**. This defines the set of all morphisms with codomain E, and ranging over all E thus defines the set of all morphisms in **Fact**.

(3) The composition of morphisms is defined as follows. Suppose we have morphisms of finite sets $\alpha : \underline{n} \to \underline{k}$ and $\beta : \underline{m} \to \underline{n}$ and an expression $E = [p_1 p_2 ... p_k]$. Write
$$E^\alpha = [p_{\alpha(1)} p_{\alpha(2)} ... p_{\alpha(n)}] = [q_1 ... q_n].$$
Let $\gamma = \alpha \circ \beta : \underline{m} \to \underline{k}$. Then the composition of the morphisms
$$\alpha_E : E^\alpha = [p_{\alpha(1)} p_{\alpha(2)} ... p_{\alpha(n)}] \to E = [p_1 ... p_k], \quad \beta_{E^\alpha} : (E^\alpha)^\beta = [q_{\beta(1)} ... q_{\beta(m)}] \to E^\alpha = [q_1 ... q_n]$$
is given by $\gamma_E : E^\gamma \to E$, where we use that $[q_{\beta(1)} ... q_{\beta(m)}] = [p_{\gamma(1)} ... p_{\gamma(m)}]$.

It's easy to check that the composition of morphisms in **Fact** is associative and has identities, so Definition 3.4 really does define a category.

Notation 3.5. Let Σ_k denote the symmetric group on k letters. Given an element $\varphi \in \Sigma_k$, we also denote the associated automorphism of $\underline{k}$ by φ.

Example 3.6. Suppose $E = [222]$. Then every permutation $\varphi \in \Sigma_3$ defines an automorphism $E^\varphi = E \to E$ in **Fact**. Conversely, every automorphism of $[222]$ uniquely corresponds to an element of Σ_3.

Suppose $E = [232]$. Then the transposition $\sigma = (13) \in \Sigma_3$ defines an automorphism of E since $E^\sigma = E$. On the other hand, the transposition $\tau = (12) \in \Sigma_3$ defines a morphism $E^\tau = [322] \to E = [232]$.

Remark 3.7. Let **FinSet** denote the category of finite sets (or rather a skeleton, with objects given by the sets $\underline{n}$ for $n \geq 0$). Given an object $\underline{k} \in$ **FinSet**, let $\mathbf{FinSet}^{/\underline{k}}$ denote the overcategory, whose objects are morphisms $[\alpha : \underline{n} \to \underline{k}]$ and whose morphisms are commuting triangles. Recall that this category has a final object given by $[\mathrm{id}_{\underline{k}}]$.

Then for every expression $E = [p_1 ... p_k]$ of length k, we have a functor
$$F_E : \mathbf{FinSet}^{/\underline{k}} \to \mathbf{Fact}$$
that sends the object $[\alpha : \underline{n} \to \underline{k}]$ to E^α and the unique morphism $[\alpha] \to [\mathrm{id}_{\underline{k}}]$ to $\alpha_E : E^\alpha \to E$. This functor has every morphism in **Fact** with codomain E in its image.

Remark 3.8. In fact, we can identify **Fact** itself as a certain overcategory (or rather, a full subcategory thereof). Namely, let $\mathcal{P}$ denote the infinite set of primes $\{2, 3, 5, ...\}$, let **Set** be the category of sets, and let $\mathbf{FinSet}^{/\mathcal{P}}$ be the full subcategory of $\mathbf{Set}^{/\mathcal{P}}$ on those morphisms $X \to \mathcal{P}$ where X is a finite set. Then we have an equivalence of categories
$$\mathbf{Fact} \simeq \mathbf{FinSet}^{/\mathcal{P}}$$
that sends an expression $E = [p_1 ... p_k]$ to the morphism $E_\bullet : \underline{k} \to \mathcal{P}$ given by $i \mapsto p_i$. Under this equivalence, the functor F_E of Remark 3.7 identifies with the functor
$$\mathbf{FinSet}^{/\underline{k}} \simeq (\mathbf{FinSet}^{/\mathcal{P}})^{/E_\bullet} \to \mathbf{FinSet}^{/\mathcal{P}}$$
that forgets the map to $E_\bullet$.

We now explain how to associate a layout to every morphism in **Fact**.

Definition 3.9. Suppose $E = [p_1 ... p_k]$ and $\alpha : \underline{n} \to \underline{k}$. We define a layout $L_{(E,\alpha)}$ as follows:[9]

(1) Its shape tuple is $(p_{\alpha(1)}, p_{\alpha(2)}, ..., p_{\alpha(n)})$.

(2) Its stride tuple is $(d_1, d_2, ..., d_n)$ where $d_i = \prod_{j<\alpha(i)} p_j$.[10]

We also let $f_{(E,\alpha)}$ denote the associated layout function.

Example 3.10. Suppose $E = [23]$ and $\varphi = (12) \in \Sigma_2$ is the nontrivial transposition. Then $L_{(E,\varphi)} = (3, 2) : (2, 1)$.

Suppose $E = (222)$ and $\varphi = (231) \in \Sigma_3$, so φ is a cycle of order 3 with $\varphi(1) = 2, \varphi(2) = 3, \varphi(3) = 1$. Then $L_{(E,\varphi)} = (2, 2, 2) : (2, 4, 1)$.

Remark 3.11. Let $E = [p_1 ... p_k]$ and $\alpha : \underline{n} \to \underline{k}$. Let $N = p_1 \cdot ... \cdot p_k$ and $N^\alpha = p_{\alpha(1)} \cdot ... \cdot p_{\alpha(n)}$. In what follows, consider the canonical isomorphisms

$$[0, N) \cong [0, p_1) \times [0, p_2) \times ... \times [0, p_k),$$
$$[0, N^\alpha) \cong [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$$

Then the associated layout function $f_{(E,\alpha)} : [0, N^\alpha) \to [0, N) \subset \mathbb{N}$ can be described as the multilinear function

$$[0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)}) \to [0, p_1) \times [0, p_2) \times ... \times [0, p_k)$$

that sends the basis vector δ_i for $1 \le i \le n$ to $\delta_{\alpha(i)}$, and which restricts to an isomorphism $[0, p_{\alpha(i)}) \xrightarrow{\cong} [0, p_{\alpha(i)})$ for all $1 \le i \le n$. In particular, if α is itself a bijection, then $f_{(E,\alpha)}$ restricts to an automorphism of $[0, N)$.

Elaborating on Remark 3.11, we have the following lemma, which indicates that composition in the category **Fact** is compatible with the composition of layout functions.

Lemma 3.12. *Suppose we have morphisms of finite sets $\alpha : \underline{n} \to \underline{k}$, $\beta : \underline{m} \to \underline{n}$ and an expression $E = [p_1 p_2 ... p_k]$. Write $\gamma = \alpha \circ \beta$. Consider the composition*

$$\gamma_E : E^\gamma = (E^\alpha)^\beta \xrightarrow{\beta_{E^\alpha}} E^\alpha \xrightarrow{\alpha_E} E$$

in **Fact**. *Then the associated layout functions satisfy the composition equality*

$$f_{(E,\gamma)} = f_{(E,\alpha)} \circ f_{(E^\alpha,\beta)}.$$

Proof. Let $N = p_1 \cdot ... \cdot p_k$, $N^\alpha = p_{\alpha(1)} \cdot ... \cdot p_{\alpha(k)}$, and $N^\gamma = p_{\gamma(1)} \cdot ... \cdot p_{\gamma(m)}$. We use the canonical isomorphisms

$$[0, N) \cong [0, p_1) \times [0, p_2) \times ... \times [0, p_k),$$
$$[0, N^\alpha) \cong [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$$
$$[0, N^\gamma) \cong [0, p_{\gamma(1)}) \times [0, p_{\gamma(2)}) \times ... \times [0, p_{\gamma(m)})$$

to write the domains and codomains of the layout functions in question (noting that $f_{(E^\alpha,\beta)}$ has codomain lying inside $[0, N^\alpha)$). We are trying to equate the multilinear function

$$f_{(E,\gamma)} : [0, p_{\gamma(1)}) \times [0, p_{\gamma(2)}) \times ... \times [0, p_{\gamma(m)}) \to [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$$

with the composition of the two multilinear functions

$$f_{(E^\alpha,\beta)} : [0, p_{\gamma(1)}) \times [0, p_{\gamma(2)}) \times ... \times [0, p_{\gamma(m)}) \to [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$$
$$f_{(E,\alpha)} : [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)}) \to [0, p_1) \times [0, p_2) \times ... \times [0, p_k).$$

But since basis vectors are mapped to basis vectors by Remark 3.11, it suffices to check the desired equality on basis vectors, which is straightforward. □

[9] If $n = 0$, then we let $L_{(E,\alpha)}$ be the "trivial layout" $(1) : (1)$.

[10] In particular, $d_i = 1$ if $\alpha(i) = 1$.

Warning 3.13. In Lemma 3.12, the per-mode condition of admissibility for composition (Definition 2.12) is obviously satisfied. However, the disjointness condition in Definition 2.17 may be violated in the case where $\beta : \underline{m} \to \underline{n}$ is not an injective function. This isn't a contradiction with the prior analysis carried out in the proof of Theorem 2.18, since there we were concerned with the composition being well-defined in the situation of *arbitrary* strides for the second layout.

We now define a "realization" functor from **Fact** to **FinSet** that sends morphisms of ordered factorizations to their associated layout functions.

Definition 3.14. Let $R : \mathbf{Fact} \to \mathbf{FinSet}$ be the functor defined as follows:

(1) Let $E = [p_1 ... p_k]$ be an object of **Fact** and let $N = p_1 \cdot ... \cdot p_k$. Then $R(E) = [0, N)$.[11]

(2) For every morphism $\alpha_E : E^\alpha \to E$, let $R(\alpha_E) = f_{(E,\alpha)} : [0, N^\alpha) \to [0, N)$ be as in Definition 3.9.

By Lemma 3.12, $R : \mathbf{Fact} \to \mathbf{FinSet}$ does indeed define a functor since it respects the composition of morphisms (and identities as well, obviously).

We note that R doesn't contain every possible function expressible as a layout function in its image. However, it does contain every automorphism $[0, N) \xrightarrow{\cong} [0, N)$ expressible as a layout function in its image.

Proposition 3.15. *Let $N > 0$ be a positive integer and let $f : [0, N) \to [0, N)$ be an automorphism such that there exists a layout L of size N with $f = f_L$.[12] Then f_L is in the image of the realization functor R.*

Proof. Without loss of generality, we may suppose that the shape tuple of L is given by $(p_1, p_2, ..., p_k)$ where the p_i are all prime numbers and $N = p_1 \cdot ... \cdot p_k$.[13] So we may write $L = (p_1, p_2, ..., p_k) : (d_1, d_2, ..., d_k)$. Then the sort of L must be of the form

$$L^\varphi \coloneqq (p_{\varphi(1)}, p_{\varphi(2)}, ..., p_{\varphi(k)}) : (1, p_{\varphi(1)}, p_{\varphi(1)}p_{\varphi(2)}, ..., \Pi_{1 \le i < k} p_{\varphi(i)})$$

for some permutation $\varphi \in \Sigma_k$, in order for f_L to be an automorphism of $[0, N)$. But this means that if we let $\psi = \varphi^{-1}$ be the inverse permutation, then

$$\psi_E : E^\psi = [p_1 p_2 ... p_k] = [p_{\psi(\varphi(1))} p_{\psi(\varphi(2))} ... p_{\psi(\varphi(k))}] \to E = [p_{\varphi(1)} p_{\varphi(2)} ... p_{\varphi(k)}]$$

is a morphism in **Fact** such that $R(\psi_E) = f_L = f$. □

Remark 3.16. One way to interpret Proposition 3.15 is that if we take the maximal subgroupoid $\mathbf{Fact}^\simeq$ inside **Fact** (i.e., the subcategory on all invertible morphisms), then

$$R : \mathbf{Fact}^\simeq \to \mathbf{FinSet}$$

carves out exactly those permutations expressible as layouts. Our motivation for this description is that for a fixed integer $N > 0$, the subset Σ_N^L of Σ_N on those automorphisms expressible as layout functions is typically not a subgroup (being not generally closed under the group multiplication, i.e. composition). Instead, if we let

$$\mathbf{Fact}_N^\simeq \subset \mathbf{Fact}^\simeq$$

be the full subgroupoid on those objects $[p_1 ... p_k]$ with $N = p_1 \cdot ... \cdot p_k$, then Σ_N^L consists of those morphisms in the image of R on $\mathbf{Fact}_N^\simeq$. However, we see that Σ_N^L is closed under the operation of taking the group inverse. Moreover, in the special case that N is a prime power p^k, then Σ_N^L is in fact a subgroup and is isomorphic to Σ_k. This corresponds to $\mathbf{Fact}_{p^k}^\simeq$ being a groupoid with one object $[pp...p]$, i.e., a group.

[11] If $E = [\,]$, this means that $R(E) = [0, 1) = \{0\}$.

[12] A priori, the codomain of f_L is $\mathbb{N}$, so part of this assertion is that f_L restricts to an automorphism of $[0, N)$.

[13] The point is that we may always maximally "uncoalesce" a layout through factoring integers appearing in the shape tuple and then inserting strides as appropriate to match the layout functions.

REFERENCES

[1] *CUTLASS — CUDA Templates for Linear Algebra Subroutines*. https://github.com/NVIDIA/cutlass.
[2] *CuTe Layout Operations*. https://github.com/NVIDIA/cutlass/blob/main/media/docs/cute/02_layout_operations.md.