CuTe Layout Algebra

Introduction

CuTe layout algebra is extremely important for understanding and applying CUTLASS for accelerated computing. Despite the fact that CuTe has a documentation for its layout algebra, it cannot be understood completely without first understanding its mathematical foundations. I tried to create some proofs for the CuTe layout algebra on my own and realized that it was a huge amount of work. Gratefully, Jay Shah has created a paper "A Note on the Algebra of CuTe Layouts" that completes the CuTe layout algebra mathematical foundations that I wanted to create.

As my proofreading, I found Jay Shah's paper mostly error-free, except for a few very minor oversights and typos. However, it does skip some details without which the paper is a little bit hard to understand. In this article, based on Jay Shah's paper, I would like to provide more proofs and explanations of the CuTe layout algebra, some of which are not present in Jay Shah's paper. Most of the definitions and annotations will follow Jay Shah's paper.

This article can be read as a complement to Jay Shah's paper, but it's also completely standalone for understanding the CuTe layout algebra.

Layout Algebra Preliminaries

Definition 2.1: Layout

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

A flattened layout means that there is no internal parentheses in the shape and stride. For example, $L = (5, 2, 2) : (16, 80, 4)$ is a flattened layout, whereas $L = (5, (2, 2)) : (16, (80, 4))$ is not. Flattening a layout will not change the semantics and operations of the layout.

Definition 2.2: Layout Size, Length, and Mode

Let $\alpha \geq 0$ be an integer and $L = \mathbf{S} : \mathbf{D} = (M_0, M_1, \dots, M_\alpha) : (d_0, d_1, \dots, d_\alpha)$ be a layout. Then:

- The size of L is the product $M = M_0 \cdot M_1 \cdot \dots \cdot M_\alpha$.
- 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.

Concatenation

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 $(\mathbf{D}, \mathbf{D}')$ respectively. Then the concatenation of L and L' is given by the layout

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

and we say that (L, L') is decomposed by L and L'.

Inductively, given layouts $L_0, L_1, \dots, L_N$, we can then form the concatenation $(L_0, L_1, \dots, L_N)$. Conversely, given L a layout, L is maximally decomposed by its modes.

Isomorphism

Let $\mathbf{S} = (M_0, M_1, ..., M_\alpha)$ and $\mathbf{D} = (d_0, d_1, ..., d_\alpha)$ be the respective shape and stride tuples of $L = \mathbf{S} : \mathbf{D}$. 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 $0, 1, 2, ..., M - 1$. Then we have an isomorphism

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

Given any $x \in [0, M)$, the isomorphism ι maps x to the tuple

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

The isomorphism mapping is bijective. In our case, given any tuple $(x_0, x_1, ..., x_\alpha) \in [0, M_0) \times [0, M_1) \times ... \times [0, M_\alpha)$, the isomorphism inverse maps the tuple to the integer

$$(x_0, x_1, ..., x_\alpha) \mapsto x_0 + x_1 \cdot M_0 + x_2 \cdot M_0 \cdot M_1 + ... + x_\alpha \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}$$

It' s straightforward to verify the above isomorphism mapping is valid and proof the above isomorphism mapping is bijective (by contradiction).

One could imagine the isomorphism as a mapping between a one-dimensional coordinate and a multi-dimensional coordinate.

Definition 2.3: Layout Function

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

$$[0, M) \cong [0, M_0) \times [0, M_1) \times ... \times [0, M_\alpha) \subset \mathbb{N}^{\times(\alpha+1)} \overset{\cdot d_0, \cdot d_1, ..., \cdot d_\alpha}{\to} \mathbb{N}^{\times(\alpha+1)} \overset{+}{\to} \mathbb{N}$$

In other words, f_L is the composition of the multilinear function

$$\begin{aligned} [0, M_0) \times [0, M_1) \times ... \times [0, M_\alpha) &\to \mathbb{N} \\ (x_0, x_1, ..., x_\alpha) &\mapsto x_0 \cdot d_0 + x_1 \cdot d_1 + ... + x_\alpha \cdot d_\alpha \end{aligned}$$

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

Computing the value of a layout function f_L at a point $x \in [0, M)$ can be decomposed into computing the sum of the values of the layout function at multiple points. This is sometimes useful for computing the value of the layout function at a point handily.

Given a layout $L = (M_0, M_1, ..., M_\alpha) : (d_0, d_1, ..., d_\alpha)$ and $x \in [0, M)$,

$$x \mapsto (x_0, x_1, ..., x_\alpha) \mapsto x_0 \cdot d_0 + x_1 \cdot d_1 + ... + x_\alpha \cdot d_\alpha$$

We also have

$$\begin{aligned} x'_0 &\mapsto (x_0, 0, 0, ..., 0) \mapsto x_0 \cdot d_0 \\ x'_1 &\mapsto (0, x_1, 0, ..., 0) \mapsto x_1 \cdot d_1 \\ &\vdots \\ x'_\alpha &\mapsto (0, 0, 0, ..., x_\alpha) \mapsto x_\alpha \cdot d_\alpha \end{aligned}$$

Therefore, we have

$$f_L(x) = f_L(x'_0) + f_L(x'_1) + ... + f_L(x'_\alpha)$$

where

$$\begin{array}{rl} x'_0 & = x \mod M_0 \\ x'_1 & = \left\lfloor \frac{x}{M_0} \right\rfloor \mod M_1 \cdot M_0 \\ & \vdots \\ x'_\alpha & = \left\lfloor \frac{x}{M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}} \right\rfloor \mod M_\alpha \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1} \end{array}$$

For example, given a layout $L = (3, 2) : (2, 3)$ and $x = 5$, we have

$$\begin{aligned} f_L(5) &= f_L(5 \mod 3) + f_L\left(\left\lfloor \tfrac{5}{3} \right\rfloor \mod 2 \cdot 3\right) \\ &= f_L(2) + f_L(3) \\ &= 2 \cdot 2 + \left\lfloor \tfrac{3}{3} \right\rfloor \cdot 3 \\ &= 4 + 3 \\ &= 7 \end{aligned}$$

Extension of Layout Function

Based on the definition of layout function, the extension of the layout function f_L is the function, $\widehat{f}_L : \mathbb{N} \to \mathbb{N}$, defined by replacing M_α with ∞ in the definition of f_L, i.e., the composite

$$\mathbb{N} \cong [0, M_0) \times [0, M_1) \times ... \times [0, M_{\alpha-1}) \times \mathbb{N} \subset \mathbb{N}^{\times(\alpha+1)} \overset{\cdot d_0, \cdot d_1, ..., \cdot d_\alpha}{\to} \mathbb{N}^{\times(\alpha+1)} \overset{+}{\to} \mathbb{N}$$

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

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

The extension of the isomorphism mapping is also bijective. The inverse mapping of the extension of the isomorphism is also given by

$$(x_0, x_1, ..., x_{\alpha-1}, x_\alpha) \mapsto x_0 + x_1 \cdot M_0 + x_2 \cdot M_0 \cdot M_1 + ... + x_\alpha \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}$$

One could imagine the extension of the isomorphism defines the last dimension of the shape to be a "batch" dimension and the batch size can be infinite.

Coalescence

Coalescence simplifies the layout and does not change the layout function.

Coalescence Rules

Considering a layout with just two integral modes, $A = (N_0, N_1) : (d_0, d_1)$, we have four cases to consider:

1. $N_1 = 1$.
2. $N_0 = 1$.
3. $d_1 = N_0 d_0$.
4. Anything else.

In the first case, obviously, $A = (N_0, 1) : (d_0, d_1) = (N_0) : (d_0)$. This can be further flattened to $A = N_0 : d_0$.

In the second case, also obviously, $A = (1, N_1) : (d_0, d_1) = (N_1) : (d_1)$. This can be further flattened to $A = N_1 : d_1$.

In the third case, we have $A = (N_0, N_1) : (d_0, N_0 d_0) = (N_0 N_1) : (d_0)$. This can be further flattened to $A = N_0 N_1 : d_0$.

In the fourth case, we could do nothing and A remains the same.

There is one case that can often be misunderstood, that is $d_0 = N_1 d_1$. In this case, we have $A = (N_0, N_1) : (N_1 d_1, d_1)$. At first glance, it seems that we could coalesce A to $(N_0 N_1) : (d_1)$. However, this is not correct, because it changes the layout function.

Considering a layout with more than two integral modes, we could apply the above rules recursively, each time we could try to coalesce two adjacent integral modes, until no more coalescence is possible. This guarantees that the laytout function remains the same.

Proof

Let $L = (N_0, N_1, \dots, N_\alpha) : (d_0, d_1, \dots, d_\alpha)$ be a layout of the concatenation of layouts L_0, L_1, ..., L_α, where each $L_k = (N_k) : (d_k)$ for $k \in [0, \alpha]$. Given a coordinate $x \mapsto (x_0, x_1, \dots, x_\alpha)$, we have the layout function for layout L as follows:

$$f_L(x) \quad = x_0 d_0 + x_1 d_1 + x_2 d_2 + \dots + x_\alpha d_\alpha$$

The layout function for layout L is just the sum of the layout functions of each L_k, i.e.,

$$\begin{aligned} f_L(x) \quad &= f_{L_0}(x_0) + f_{L_1}(x_1) + \dots + f_{L_\alpha}(x_\alpha) \\ &= x_0 d_0 + x_1 d_1 + x_2 d_2 + \dots + x_\alpha d_\alpha \end{aligned}$$

Suppose two adjacent integral modes, say $A = (L_i, L_{i+1}) = (N_i, N_{i+1}) : (d_i, d_{i+1})$, can be coalesced to a new layout A_i' whose shape is $(N_i N_{i+1})$, which satisfies the first, second, and third coalescence rules. By the definition of coalescence, we must have $f_{A'}(x) = f_A(x)$.

For any $x_i \in [0, N_i)$ and $x_{i+1} \in [0, N_{i+1})$, we have $x' = x_i + x_{i+1} \cdot N_i$, where $x' \in [0, N_i N_{i+1})$.

$$\begin{aligned} f_{A'}(x') \qquad &= f_A(x') \\ &= x_i d_i + x_{i+1} d_{i+1} \\ &= f_{L_i}(x_i) + f_{L_{i+1}}(x_{i+1}) \end{aligned}$$

Given any coordinate $x \mapsto (x_0, x_1, \dots, x_{i-1}, x_i, x_{i+1}, x_{i+2}, \dots, x_\alpha)$ for the layout L, after coalescing A to A', the coordiante $x \mapsto (x_0, x_1, \dots, x_{i-1}, x', x_{i+2}, \dots, x_\alpha)$ for the layout L', where L' is the layout after coalescing A to A'.

This is easy to verify because of the isomorphism of coordinates.

$$
\begin{aligned}
x &= x_0 + x_1 \cdot N_0 + x_2 \cdot N_0 N_1 + ... + x_{i-1} \cdot N_0 N_1 ... N_{i-2} + x_i \cdot N_0 N_1 ... N_{i-1} + x_{i+1} \cdot N_0 N_1 ... N_{i-1} N_i + x \\
&= x_0 + x_1 \cdot N_0 + x_2 \cdot N_0 N_1 + ... + x_{i-1} \cdot N_0 N_1 ... N_{i-2} + (x_i + x_{i+1} \cdot N_i) \cdot N_0 N_1 ... N_{i-1} + x_{i+2} \cdot N \\
&= x_0 + x_1 \cdot N_0 + x_2 \cdot N_0 N_1 + ... + x_{i-1} \cdot N_0 N_1 ... N_{i-2} + x' \cdot N_0 N_1 ... N_{i-1} + x_{i+2} \cdot N_0 N_1 ...
\end{aligned}
$$

Then we have

$$
\begin{aligned}
f_{L'}(x) &= f_{L_0}(x_0) + f_{L_1}(x_1) + ... + f_{L_{i-1}}(x_{i-1}) + f_{A'}(x') + f_{L_{i+2}}(x_{i+2}) + ... + f_{L_\alpha}(x_\alpha) \\
&= f_{L_0}(x_0) + f_{L_1}(x_1) + ... + f_{L_{i-1}}(x_{i-1}) + f_{L_i}(x_i) + f_{L_{i+1}}(x_{i+1}) + f_{L_{i+2}}(x_{i+2}) + ... + f_{L_\alpha}(x_\alpha) \\
&= f_L(x)
\end{aligned}
$$

Therefore, the layout function remains the same after coalescing A to A'.

This concludes the proof.

By-Mode Coalescence

In some cases, when the modes are not completely integral in the layout and we would like to keep the number of modes unchanged, we could perform by-mode coalescence. This can be achieved by disabling the coalescence of adjacent integral modes from two different modes in the coalescence rules for any layout with more than two integral modes.

For example, if we have a layout of two modes $A = ((N_0, N_1, ..., N_\alpha), (N_{\alpha+1}, N_{\alpha+2}, ..., N_\beta) : ((d_0, d_1, ..., d_\alpha), (d_{\alpha+1}, d_{\alpha+2}, ..., d_\beta)))$, to perform by-mode coalescence, we could coalesce the integral modes $(N_0, N_1, ..., N_\alpha)$ and $(N_{\alpha+1}, N_{\alpha+2}, ..., N_\beta)$ separately until no more coalescence is possible for each one, and no more coalescence will be performed further between the two consequent coalesced modes even if they can be coalesced. This will result in a layout with the same number of modes as before.

Implication of Coalescence

Coalescence simplifies the layout and does not change the layout function. It is a useful operation to reduce the complexity of the layout and the related computation while preserving its functionality. If non-by-mode coalescence is performed on a layout, the layout can be simplified such that each mode is an integral mode, i.e., $A = (N_0, N_1, ..., N_\alpha) : (d_0, d_1, ..., d_\alpha)$ where N_k is an integer, not a tuple of integers, and $N_k > 1$ for $k \in [0, \alpha]$.

The property of coalescence is very important and most of the proofs we will present in the article assumes that the layout is coalesced. This assumption is fine mathematically because the coalescence does not change the layout function.

Complementation

Definition 2.4: Sorted Layout

Let $A = (N_0, N_1, ..., N_\alpha) : (d_0, d_1, ..., d_\alpha)$ be a layout. We say that A is sorted if $d_0 \leq d_1 \leq ... \leq d_\alpha$ and for every $i < j$, if $d_i = d_j$, then $N_i \leq N_j$.

Note that sorting a layout, or more generally, changing the order of modes of a layout, will change the semantics and operations of the layout.

For example, suppose we have a layout $A = (2, 4) : (4, 1)$ and a layout $B = (4, 2) : (1, 4)$. We could see that B is the sorted version of A. We could compute the layout function of A and B as follows using lookup tables:

$$
\begin{aligned}
f_A(0) &= f_A(0,0) = 0 \cdot 4 + 0 \cdot 1 = 0 \\
f_A(1) &= f_A(1,0) = 1 \cdot 4 + 0 \cdot 1 = 4 \\
f_A(2) &= f_A(0,1) = 0 \cdot 4 + 1 \cdot 1 = 1 \\
f_A(3) &= f_A(1,1) = 1 \cdot 4 + 1 \cdot 1 = 5 \\
f_A(4) &= f_A(0,2) = 0 \cdot 4 + 2 \cdot 1 = 2 \\
f_A(5) &= f_A(1,2) = 1 \cdot 4 + 2 \cdot 1 = 6 \\
f_A(6) &= f_A(0,3) = 0 \cdot 4 + 3 \cdot 1 = 3 \\
f_A(7) &= f_A(1,3) = 1 \cdot 4 + 3 \cdot 1 = 7
\end{aligned}
$$

$$
\begin{aligned}
f_B(0) &= f_B(0,0) = 0 \cdot 1 + 0 \cdot 4 = 0 \\
f_B(1) &= f_B(1,0) = 1 \cdot 1 + 0 \cdot 4 = 1 \\
f_B(2) &= f_B(2,0) = 2 \cdot 1 + 0 \cdot 4 = 2 \\
f_B(3) &= f_B(3,0) = 3 \cdot 1 + 0 \cdot 4 = 3 \\
f_B(4) &= f_B(0,1) = 0 \cdot 1 + 1 \cdot 4 = 4 \\
f_B(5) &= f_B(1,1) = 1 \cdot 1 + 1 \cdot 4 = 5 \\
f_B(6) &= f_B(2,1) = 2 \cdot 1 + 1 \cdot 4 = 6 \\
f_B(7) &= f_B(3,1) = 3 \cdot 1 + 1 \cdot 4 = 7
\end{aligned}
$$

We could see that the layout B is typically referred as the column-major layout, and the layout A is typically referred as the row-major layout. They are completely different layouts.

More generally, the sorted layout is a just like the "generalization" of the column-major layout.

Definition 2.5: Admission for Complementation

Let $A = (N_0, N_1, \dots, N_\alpha) : (d_0, d_1, \dots, d_\alpha)$ be a layout and M be a positive integer. If A is not sorted then replace A with its sorted version. We say that the pair $\{A, M\}$ is admissible for complementation (or simply admissible) if:

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

That $\{A, M\}$ is admissible for complementation also implies:

- For all $1 \leq i \leq \alpha$, $N_{i-1} \cdot d_{i-1} \leq d_i$ and $d_{i-1} \leq d_i$.
- $N_\alpha \cdot d_\alpha \leq M$ and $d_\alpha \leq M$.

Definition 2.6: Complementation

Let $A = (N_0, N_1, \dots, N_\alpha) : (d_0, d_1, \dots, d_\alpha)$ be a layout and M be a positive integer. If $\{A, M\}$ is admissible for complementation, then if A is not sorted, replace A with its sorted version. The complement of $\{A, M\}$ is defined to be the layout

$$
\text{complement}(A, M) = \left(d_0, \tfrac{d_1}{N_0 d_0}, \tfrac{d_2}{N_1 d_1}, \dots, \tfrac{d_\alpha}{N_{\alpha-1} d_{\alpha-1}}, \tfrac{M}{N_\alpha d_\alpha}\right) : (1, N_0 d_0, N_1 d_1, \dots, N_\alpha d_\alpha)
$$

Note that the size of the complement of $\{A, M\}$, $\text{size}(\text{complement}(A, M))$, is $\frac{M}{\text{size}(A)} = \frac{M}{N_0 \cdot N_1 \cdot \dots \cdot N_\alpha}$.

By definition, the complement of $\{A, M\}$ is insensitive to the order of the modes of A, since it will always be sorted before complementation.

The complement of $\{A, M\}$ is strictly increasing. This might not be very obvious, so we will show a proof.

Proof

Suppose $B = \text{complement}(A, M)$, to show that the layout function f_B, whose domain is a set of natural numbers, is strictly increasing, we need to show that for every two adjacent natural numbers x and $x+1$, $0 \leq x < x+1 < \text{size}(B)$, we have $f_B(x) < f_B(x+1)$.

Because of the isomorphism, suppose the mapping of x is as follows:

$$x \mapsto (x_0, x_1, \ldots, x_\alpha, x_{\alpha+1})$$

By definition of the layout function f_B, we have

$$f_B(x) = x_0 + x_1 \cdot N_0 d_0 + x_2 \cdot N_1 d_1 + \ldots + x_\alpha \cdot N_{\alpha-1} d_{\alpha-1} + x_{\alpha+1} \cdot N_\alpha d_\alpha$$

The mapping of $x+1$ can have many different cases.

In the simplest case,

$$x+1 \mapsto (x_0 + 1, x_1, \ldots, x_\alpha, x_{\alpha+1})$$

Then we have

$$\begin{aligned} f_B(x+1) &= x_0 + 1 + x_1 \cdot N_0 d_0 + x_2 \cdot N_1 d_1 + \ldots + x_\alpha \cdot N_{\alpha-1} d_{\alpha-1} + x_{\alpha+1} \cdot N_\alpha d_\alpha \\ &= f_B(x) + 1 \\ &> f_B(x) \end{aligned}$$

In a more complicated case, where $x_0 = d_0 - 1$ and $x_1 < \frac{d_1}{N_0 d_0} - 1$, we have

$$x+1 \mapsto (0, x_1 + 1, \ldots, x_\alpha, x_{\alpha+1})$$

Then we have

$$\begin{aligned} f_B(x+1) &= 0 + (x_1 + 1) \cdot N_0 d_0 + x_2 \cdot N_1 d_1 + \ldots + x_\alpha \cdot N_{\alpha-1} d_{\alpha-1} + x_{\alpha+1} \cdot N_\alpha d_\alpha \\ &= f_B(x) - x_0 + N_0 d_0 \\ &= f_B(x) - (d_0 - 1) + N_0 d_0 \\ &= f_B(x) + 1 + (N_0 - 1) d_0 \\ &> f_B(x) \end{aligned}$$

Because $N_0 \geq 1$, we have $(N_0 - 1) d_0 \geq 0$, so we have

$$f_B(x+1) > f_B(x)$$

In general, when $x_0 = d_0 - 1$, for some $k \in [1, \alpha - 1]$, $x_i = \frac{d_i}{N_{i-1} d_{i-1}} - 1$ for every $i \in [1, k]$, $x_{k+1} < \frac{d_{k+1}}{N_k d_k} - 1$, we have

$$x+1 \mapsto (0, 0, \ldots, 0, x_{k+1} + 1, \ldots, x_\alpha, x_{\alpha+1})$$

Then we have

$$
\begin{aligned}
f_B(x+1) &= 0 + 0 \cdot N_0 d_0 + \ldots + 0 \cdot N_{k-1} d_{k-1} + (x_{k+1}+1) \cdot N_k d_k + \ldots + x_\alpha \cdot N_{\alpha-1} d_{\alpha-1} + x_{\alpha+1} \cdot N_\alpha d_\alpha \\
&= f_B(x) - x_0 - \left(\sum_{i=1}^{k} x_i \cdot N_{i-1} d_{i-1} \right) + N_k d_k \\
&= f_B(x) - (d_0 - 1) - \left(\sum_{i=1}^{k} \left(\frac{d_i}{N_{i-1} d_{i-1}} - 1 \right) \cdot N_{i-1} d_{i-1} \right) + N_k d_k \\
&= f_B(x) - (d_0 - 1) - \left(\sum_{i=1}^{k} \left(d_i - N_{i-1} d_{i-1} \right) \right) + N_k d_k \\
&= f_B(x) - (d_0 - 1) + \sum_{i=1}^{k} N_{i-1} d_{i-1} - \sum_{i=1}^{k} d_i + N_k d_k \\
&= f_B(x) + \sum_{i=0}^{k} \left(N_i - 1 \right) d_i + 1
\end{aligned}
$$

Because $N_i \geq 1$ for every i, we have $(N_i - 1)\, d_i \geq 0$ for every i, so we have

$$
f_B(x+1) \quad > f_B(x)
$$

This concludes the proof.

Similarly, we could also prove that the extension of the complement of $\{A, M\}$ is strictly increasing.

Proposition 2.7

Let $\{A = (N_0, N_1, \ldots, N_\alpha) : (d_0, d_1, \ldots, d_\alpha), M\}$ be admissible for complementation 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

Because $\text{size}(A) = \prod_{i=0}^{\alpha} N_i$ and $\text{size}(B) = \frac{M}{\prod_{i=0}^{\alpha} N_i}$, we have $\text{size}(C) = \text{size}(A) \cdot \text{size}(B) = M$. Thus the domain of f_C is $[0, M)$.

Note that the image of f_C is the same as that of $f_{C'}$ for any permutation C' of C.

To see this, suppose we have the following layout C and its permutation C' in which only one pair of the modes is permuted.

$$
\begin{aligned}
C &= \left(N_0, N_1, \ldots, N_i, \ldots, N_j, \ldots, N_\alpha\right) : \left(d_0, d_1, \ldots, d_i, \ldots, d_j, \ldots, d_\alpha\right) \\
C' &= \left(N_0, N_1, \ldots, N_j, \ldots, N_i, \ldots, N_\alpha\right) : \left(d_0, d_1, \ldots, d_j, \ldots, d_i, \ldots, d_\alpha\right)
\end{aligned}
$$

The domains of f_C and f'_C are both $[0, M)$. For any $x_C \in [0, M)$, we have

$$
\begin{aligned}
x_C &\mapsto \left(x_0, x_1, \ldots, x_i, \ldots, x_j, \ldots, x_\alpha\right) \\
x_{C'} &\mapsto \left(x_0, x_1, \ldots, x_j, \ldots, x_i, \ldots, x_\alpha\right)
\end{aligned}
$$

and x_C and $x_{C'}$ are bijective.

Because by definition, $f_C(x_C) = f_{C'}(x_{C'})$, the image of f_C is the same as that of $f_{C'}$.

For any permutation C' of C, it can be obtained by permuting one pair of the modes of C at a time and each time the image of f_C is the same as that of $f_{C'}$. Therefore, the image of f_C is the same as that of $f_{C'}$ for any permutation C' of C.

When computing the image of f_C we may sort C. Without loss of generality, suppose $A = (N_0, N_1, \dots, N_\alpha) : (d_0, d_1, \dots, d_\alpha)$ is already sorted. After sorting C, the sorted C' could only be as follows:

$$C' = \left(d_0, N_0, \tfrac{d_1}{N_0 d_0}, N_1, \tfrac{d_2}{N_1 d_1}, N_2, \dots, \tfrac{d_\alpha}{N_{\alpha-1} d_{\alpha-1}}, N_\alpha, \tfrac{M}{N_\alpha d_\alpha}\right) : (1, d_0, N_0 d_0, d_1, N_1 d_1, d_2, \dots, N_{\alpha-1} d_{\alpha-1}, d_\alpha, N_$$

Because $d_i \le N_i d_i$ and $N_i d_i \le d_{i+1}$ for every i, when $N_i = 1$, $N_i \le \frac{d_{i+1}}{N_i d_i}$, when $N_i d_i = d_{i+1}$, $\frac{d_{i+1}}{N_i d_i} \le N_{i+1}$, thus C' is sorted and any permutation of C' will make it not sorted.

Then we may rewrite

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

where $\beta = 2\alpha + 1$ and the maximum value that $f_{C'}$ attains is computed as follows:

$$\begin{aligned} f_{C'}(M-1) &= f_{C'}(r_0 - 1, r_1 - 1, r_2 - 1, \dots, r_{\beta-1} - 1, r_\beta - 1) \\ &= (r_0 - 1) + (r_1 - 1) \cdot r_0 + (r_2 - 1) \cdot r_0 r_1 + \dots + (r_{\beta-1} - 1) \cdot r_0 r_1 \dots r_{\beta-2} + (r_\beta - 1) \cdot r_0 r_1 \dots r \\ &= r_0 - 1 + r_0 r_1 - r_0 + r_0 r_1 r_2 - r_0 r_1 + \dots + r_0 r_1 \dots r_{\beta-1} - r_0 r_1 \dots r_{\beta-2} + r_0 r_1 \dots r_\beta - r_0 r_1 \dots r \\ &= r_0 r_1 \dots r_\beta - 1 \\ &= M - 1 \end{aligned}$$

Then in this case, to establish the bijectivity assertion, it' s sufficient to just show $f_{C'}(x)$ is injective, i.e., for any $x, y \in [0, M)$, if $f_{C'}(x) = f_{C'}(y)$, then $x = y$.

Suppose the isomorphism mapping of x and y are as follows:

$$\begin{aligned} x &\mapsto (x_0, x_1, \dots, x_\beta) \\ y &\mapsto (y_0, y_1, \dots, y_\beta) \end{aligned}$$

Because $f_{C'}(x) = f_{C'}(y)$, we have

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

We will use strong induction to show that $x_i = y_i$ for every $i \in [0, \beta]$.

Because $f_{C'}(x) \mod r_0 = f_{C'}(y) \mod r_0$, we have $x_0 = y_0$.

Now suppose by the strong induction that given $i \in (0, \beta]$, for all $j < i$, we have $x_j = y_j$. we have

$$x_i \cdot r_0 r_1 \dots r_{i-1} + x_{i+1} \cdot r_0 r_1 \dots r_i + \dots + x_\beta \cdot r_0 r_1 \dots r_{\beta-1} = y_i \cdot r_0 r_1 \dots r_{i-1} + y_{i+1} \cdot r_0 r_1 \dots r_i + \dots + y_\beta \cdot r_0 r$$

Because $x_i \in [0, r_i)$ and $y_i \in [0, r_i)$, taking this equation modulo $r_0 r_1 \dots r_i$ and dividing by $r_0 r_1 \dots r_{i-1}$, we have $x_i = y_i$.

Because $(x_0, x_1, \dots, x_\beta) = (y_0, y_1, \dots, y_\beta)$, and the isomorphism mapping is bijective, we have $x = y$.

Therefore $f_{C'} : [0, M) \to \mathbb{N}$ restricts to a bijection $[0, M) \cong [0, M)$. So does f_C.

This concludes the proof.

Corollary 2.8 Complementation Disjointness

The Corollary 2.8 explains what it means of taking a complement of a layout.

In the setting of Proposition 2.7, let $I = [0, \text{size}(A)) = [0, N_0 N_1 \dots N_\alpha)$ be the domain of f_A. Then

$$f_A(I) \cap \hat{f}_B(I) = \{0\}$$

In other words, $\hat{f}_A$ and $\hat{f}_B$ have disjoint image when restricted to the domain of f_A, apart from 0.

Note that in the corollary, f_A and $\hat{f}_A$ are actually interchangeable, because the function domain is restricted to the domain of f_A.

Proof

Let $J = [0, \text{size}(B)) = [0, \frac{M}{N_0 N_1 \dots N_\alpha})$ be the domain of f_B. Then by Proposition 2.7, we have

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

To understand this, for any $x_A \in I$ and any $x_B \in J$, because of the isomorphism, we have

$$\begin{aligned} x_A &\mapsto (x_{A,0}, x_{A,1}, \dots, x_{A,\alpha}) \\ x_B &\mapsto (x_{B,0}, x_{B,1}, \dots, x_{B,\alpha}, x_{B,\alpha+1}) \end{aligned}$$

Then we have

$$\begin{aligned} f_A(x_A) &= x_{A,0} + x_{A,1} \cdot N_0 + x_{A,2} \cdot N_0 N_1 + \dots + x_{A,\alpha} \cdot N_0 N_1 \dots N_{\alpha-1} \\ f_B(x_B) &= x_{B,0} + x_{B,1} \cdot N_0 d_0 + x_{B,2} \cdot N_1 d_1 + \dots + x_{B,\alpha} \cdot N_{\alpha-1} d_{\alpha-1} + x_{B,\alpha+1} \cdot N_\alpha d_\alpha \end{aligned}$$

We orchestrate new coordinates for layout C as follows:

$$\begin{aligned} x'_A &\mapsto (0, x_{A,0}, 0, x_{A,1}, 0, x_{A,2}, \dots, 0, x_{A,\alpha}, 0) \\ x'_B &\mapsto (x_{B,0}, 0, x_{B,1}, 0, x_{B,2}, \dots, x_{B,\alpha}, 0, x_{B,\alpha+1}) \end{aligned}$$

Then we have

$$\begin{aligned} f_C(x'_A) &= x_{A,0} + x_{A,1} \cdot N_0 + x_{A,2} \cdot N_0 N_1 + \dots + x_{A,\alpha} \cdot N_0 N_1 \dots N_{\alpha-1} \\ &= f_A(x_A) \\ f_C(x'_B) &= x_{B,0} + x_{B,1} \cdot N_0 d_0 + x_{B,2} \cdot N_1 d_1 + \dots + x_{B,\alpha} \cdot N_{\alpha-1} d_{\alpha-1} + x_{B,\alpha+1} \cdot N_\alpha d_\alpha \\ &= f_B(x_B) \end{aligned}$$

By the Proposition 2.7, we have $f_C : [0, M) \to \mathbb{N}$ restricts to a bijection $[0, M) \cong [0, M)$. If $x'_A \neq x'_B$, then $f_C(x'_A) \neq f_C(x'_B)$, and $f_A(x_A) \neq f_B(x_B)$.

Obviously, other than $(0, 0, \dots, 0)$, for any values of $x_{A,0}, x_{A,1}, \dots, x_{A,\alpha}$ and $x_{B,0}, x_{B,1}, \dots, x_{B,\alpha}, x_{B,\alpha+1}$, $\big(0, x_{A,0}, 0, x_{A,1}, 0, x_{A,2}, \dots, 0, x_{A,\alpha}, 0\big) \neq \big(x_{B,0}, 0, x_{B,1}, 0, x_{B,2}, \dots, x_{B,\alpha}, 0, x_{B,\alpha+1}\big)$, $x'_A \neq x'_B$, $f_C(x'_A) \neq f_C(x'_B)$, and $f_A(x_A) \neq f_B(x_B)$.

This means, for any $x \in I$ that $x \neq 0$, there is no $y \in J$ such that $f_A(x) = f_B(y)$.

When $x = 0$, we have $f_A(x) = f_B(x) = 0$. Thus we could claim that

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

In the Definition 2.6: Complementation, we have shown that the complement of $\{A, M\}$, f_B, as well as its extension $\hat{f}_B$, are strictly increasing.

In addition, by the extension of the isomorphism, we have

$$\text{size}(B) \mapsto \left(0, 0, \dots, 0, \tfrac{M}{N_\alpha d_\alpha}\right)$$

Then we have

$$\begin{aligned} \hat{f}_B(\text{size}(B)) \quad &= 0 + 0 \cdot 1 + 0 \cdot N_0 d_0 + \dots + 0 \cdot N_{\alpha-1} d_{\alpha-1} + \tfrac{M}{N_\alpha d_\alpha} \cdot N_\alpha d_\alpha \\ &= M \end{aligned}$$

The largest value attained by f_A is at $N_0 N_1 \dots N_\alpha - 1$, and $f_A(N_0 N_1 \dots N_\alpha - 1) = (N_0 - 1)d_0 + (N_1 - 1)d_1 + \dots + (N_\alpha - 1)d_\alpha$.

Because $(N_0 - 1)d_0 < N_0 d_0$ and $N_i d_i \leq d_{i+1}$ for every $i \in [0, \alpha - 1]$, $N_\alpha d_\alpha \leq M$, we have

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

Thus $f_A(N_0 N_1 \dots N_\alpha - 1) < \hat{f}_B(\text{size}(B))$.

In the case of $I \cap J = I$, i.e., $\text{size}(A) \leq \text{size}(B)$. Then we have

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

Because in this case, $f_B(I) = \hat{f}_B(I)$, we have

$$f_A(I) \cap \hat{f}_B(I) = \{0\}$$

In the other case of $I \cap J = J$, i.e., $\text{size}(A) \geq \text{size}(B)$. Because the largest value attained by f_A is $f_A(N_0 N_1 \dots N_\alpha - 1)$, and $f_A(N_0 N_1 \dots N_\alpha - 1) < \hat{f}_B(\text{size}(B))$, for any $x \in I/J$, we have $f_A(x) < \hat{f}_B(\text{size}(B))$.

Thus,

$$f_A(I) \cap \hat{f}_B(I/J) = \emptyset$$

Therefore,

$$\begin{aligned} f_A(I) \cap \hat{f}_B(I) &= f_A(I) \cap \left(\hat{f}_B(I) \cup \hat{f}_B(I/J)\right) \\ &= f_A(I) \cap \left(f_B(I) \cup \hat{f}_B(I/J)\right) \\ &= (f_A(I) \cap f_B(I)) \cup \left(f_A(I) \cap \hat{f}_B(I/J)\right) \\ &= \{0\} \cup \emptyset \\ &= \{0\} \end{aligned}$$

Taken together, we have

$$f_A(I) \cap \hat{f}_B(I) = \{0\}$$

This concludes the proof.

A short note on the original proof of Corollary 2.8 in the paper is that Jay Shah claimed $f_A(I \cap J) \cap f_B(I \cap J) = \{0\}$, which is insufficient to show the proof. The sufficient statement should be $f_A(I) \cap f_B(J) = \{0\}$.

Remark 2.9 Complementation Disjointness, Ordering, and Boundedness

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 $f_A(x) \neq \hat{f}_B(y)$ for all $x \neq 0$ and y in the domain of f_A.
2. B is ordered in the sense that $f_B(x)$ is a strictly increasing function.
3. B is bounded by M in the sense that $\text{size}(B) \geq \frac{M}{\text{size}(A)}$ and $\text{cosize}(B) \leq \left\lfloor \frac{M}{\text{cosize}(A)} \right\rfloor \cdot \text{cosize}(A)$. Here the cosize of a layout L is defined as $\text{cosize}(L) = f_L(\text{size}(L) - 1) + 1$.

The property 1 and 2 have been proved in the Corollary 2.8 and Definition 2.6. We will show a proof of the property 3.

Proof

By Definition 2.6, we have $\text{size}(B) = \frac{M}{\text{size}(A)}$.

Because cosize is insensitive to the ordering of the layout, without loss of generality, we sorted A so that $A = (N_0, N_1, \ldots, N_\alpha) : (d_0, d_1, \ldots, d_\alpha)$.

By the definition of cosize, we have

$$\begin{aligned} \text{cosize}(B) &= f_B(\text{size}(B) - 1) + 1 \\ &= f_B\left(d_0 - 1, \tfrac{d_1}{N_0 d_0} - 1, \ldots, \tfrac{d_\alpha}{N_{\alpha-1} d_{\alpha-1}} - 1, \tfrac{M}{N_\alpha d_\alpha} - 1\right) + 1 \\ &= (d_0 - 1) + \left(\tfrac{d_1}{N_0 d_0} - 1\right) \cdot N_0 d_0 + \ldots + \left(\tfrac{d_\alpha}{N_{\alpha-1} d_{\alpha-1}} - 1\right) \cdot N_{\alpha-1} d_{\alpha-1} + \left(\tfrac{M}{N_\alpha d_\alpha} - 1\right) \cdot N_\alpha d_\alpha + 1 \\ &= d_0 + d_1 + \ldots + d_\alpha - N_0 d_0 - N_1 d_1 - \ldots - N_\alpha d_\alpha + M \\ &= M - ((N_0 - 1)\, d_0 + (N_1 - 1)\, d_1 + \ldots + (N_\alpha - 1)\, d_\alpha) \\ &= M - f_A(\text{size}(A) - 1) \\ &= M - (\text{cosize}(A) - 1) \end{aligned}$$

To obtain the inequality $\text{cosize}(B) \leq \left\lfloor \frac{M}{\text{cosize}(A)} \right\rfloor \cdot \text{cosize}(A)$, we divide the above equation by $\text{cosize}(A)$.

$$\begin{aligned} \frac{\text{cosize}(B)}{\text{cosize}(A)} &= \frac{M-(\text{cosize}(A)-1)}{\text{cosize}(A)} \\ &= \frac{M}{\text{cosize}(A)} - 1 + \frac{1}{\text{cosize}(A)} \end{aligned}$$

and we have to show that

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

In fact, for any $a, b \in \mathbb{N}$ and $a \geq 1$, we have

$$\frac{b}{a} - 1 + \frac{1}{a} \leq \left\lfloor \frac{b}{a} \right\rfloor$$

To see this, suppose $\frac{b}{a} = \left\lfloor \frac{b}{a} \right\rfloor + c$, where $c = \frac{k}{a}$ and k is an integer such that $0 \leq k < a$. Then we have
$\frac{1}{a} \leq c < 1$ and $1 \leq ac < a$. Then we want to show that

$$\frac{b}{a} - 1 + \frac{1}{a} \leq \frac{b}{a} - c$$

$$-a + 1 \leq -ac$$

$$a - ac \geq 1$$

$$a - k \geq 1$$

Because a and k are both integers and $0 \leq k < a$, we have $a - k \geq 1$. Thus the inequality holds.

This concludes the proof.

Non-Integral Layout Complementation

All the properties and proofs of complementation above assumes that the layout being complemented is a layout whose mode is integral, i.e., $A = (N_0, N_1, \ldots, N_\alpha) : (d_0, d_1, \ldots, d_\alpha)$. In the case where the layout is non-integral, i.e., some of the modes are not integers, the layout shall be coalesced to an integral layout before the complementation is applied. This is valid because coalescence does not change the layout function.

Implication of Complementation

The complementation of a layout finds a complement layout with a positive integer so that when the two layouts are concatenated, such as $(B, \text{complement}(B, M))$, the new layout is a bijection $[0, M) \cong [0, M)$. This is also saying, if the original layout is repeated using the complement layout, the new layout is still a bijection.

Composition

Definition 2.11 Left Divisibility

Let $M, d > 0$ be positive integers and let $M = M_0 \cdot M_1 \cdot ... \cdot M_\alpha$ be a given factorization of M by integers $M_k > 1$ for $k \in [0, \alpha]$. Replacing M_α by ∞, let

$$\hat{M} = M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1} \cdot \infty$$

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. $M_0 \cdot M_1 \cdot ... \cdot M_{i-1}$ divides d.
2. Suppose the first condition is satisfied. Let $c = \frac{d}{M_0 \cdot M_1 \cdot ... \cdot M_{i-1}}$. Then if $i < \alpha$, we require in addition that $1 \leq c < M_i$.
3. For the second condition in the case $i < \alpha$, we require in addition that c also divides M_i.

Here i is necessarily unique if it exists. We could prove this by contradiction.

Proof

Suppose there exists two distinct i and j such that the three conditions are satisfied. Without loss of generality, suppose $i < j$.

There are two cases to consider.

In the case where $j < \alpha$, we will also have $i < \alpha$. Then we have

$$d = c \cdot M_0 \cdot M_1 \cdot ... \cdot M_{i-1}$$

where c is some positive integer such that $1 \leq c < M_i$.

Similarly,

$$d = c' \cdot M_0 \cdot M_1 \cdot ... \cdot M_{j-1}$$

where c' is some positive integer such that $1 \leq c' < M_j$.

Thus,

$$c \cdot M_0 \cdot M_1 \cdot ... \cdot M_{i-1} = c' \cdot M_0 \cdot M_1 \cdot ... \cdot M_{j-1}$$

$$c = c' \cdot M_i \cdot M_{i+1} \cdot ... \cdot M_{j-1}$$

To make the above equation valid, we must show

$$c' \cdot \frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-1} = 1$$

However, because $M_k > 1$ for $k \in [0, \alpha]$, $\frac{M_i}{c} > 1$, and $c' \geq 1$, it is not possible to have the above equation valid. This raises a contradiction. Therefore, i is unique.

In the case where $j = \alpha$, we will also have $i < \alpha$. Then we have

$$d = c \cdot M_0 \cdot M_1 \cdot ... \cdot M_{i-1}$$

where c is some positive integer such that $1 \leq c < M_i$.

Similarly,

$$d \ = c' \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}$$

where c' is some positive integer.

Thus,

$$c \cdot M_0 \cdot M_1 \cdot ... \cdot M_{i-1} \ = c' \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}$$

$$c \ = c' \cdot M_i \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1}$$

To make the above equation valid, we must show

$$c' \cdot \tfrac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1} = 1$$

However, because $M_k > 1$ for $k \in [0, \alpha]$, $\frac{M_i}{c} > 1$, and $c' \geq 1$, it is not possible to have the above equation valid. This raises a contradiction. Therefore, i is unique.

Taken together, i is unique if it exists.

This concludes the proof.

If i exists, we will refer to i as the division index and write $\hat{M} = d \cdot \hat{M}'$, where $\hat{M}'$ is endowed with the following induced factorization:

1. If $0 \leq i < \alpha$, then $\hat{M}' = \hat{M}'_0 \cdot \hat{M}'_1 \cdot ... \cdot \hat{M}'_{\alpha-i-1} \cdot \infty$ with $\hat{M}'_0 = \frac{M_i}{c} > 1$ and $\hat{M}'_j = M_{i+j}$ for $0 < j < \alpha - i$.
2. If $i = \alpha$, then $\hat{M} = d \cdot \infty$ and we will let $\hat{M}' = \infty$.

To see this, in the case where $0 \leq i < \alpha$, we have

$$\begin{aligned}
\hat{M} &= M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1} \cdot \infty \\
&= M_0 \cdot M_1 \cdot ... \cdot M_{i-1} \cdot M_i \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1} \cdot \infty \\
&= \tfrac{d}{c} \cdot M_i \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1} \cdot \infty \\
&= d \cdot \tfrac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1} \cdot \infty \\
&= d \cdot \hat{M}'_0 \cdot \hat{M}'_1 \cdot ... \cdot \hat{M}'_{\alpha-i-1} \cdot \infty \\
&= d \cdot \hat{M}'
\end{aligned}$$

where $\hat{M}' = M'_0 \cdot M'_1 \cdot ... \cdot M'_{\alpha-i-1} \cdot \infty$ with $M'_0 = \frac{M_i}{c} > 1$ and $M'_j = M_{i+j}$ for $0 < j < \alpha - i$.

In the case where $i = \alpha$, we have

$$\begin{aligned}
\hat{M} &= M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1} \cdot \infty \\
&= \tfrac{d}{c} \cdot \infty \\
&= d \cdot \infty \\
&= d \cdot \hat{M}'
\end{aligned}$$

where $\hat{M}' = \infty$.

Furthermore, we say that M is weakly left divisible by d if there exists $0 \leq i \leq \alpha$ such that the conditions 1 and 2 are satisfied for left divisibility, but not necessarily the condition 3.

Notice that in the proof of the uniqueness of division index i, we have never used the condition 3. Therefore, we could still the necessarily unique i the division index for weak left divisibility, but we no longer have the factorization of $\hat{M}$, because the factorization assumes the condition 3 of left divisibility.

Also notice that $\hat{M}'$ with its induced factorization can itself be considered for left divisibility or weak left divisibility (with the step or replacing the last factor by ∞ now being superfluous). More specifically, because $\hat{M}' > 0$, $\hat{M}'_j > 1$ for $j \in [0, \alpha - i - 1]$, and $\hat{M}' = \hat{M}'_0 \cdot \hat{M}'_1 \cdot ... \cdot \hat{M}'_{\alpha-i-1} \cdot \infty$, given another positive integer $d' > 0$, we could completely test whether the properties of left divisibility or weak left divisibility hold for $\hat{M}'$ with respect to d'. Replacing the last factor by ∞ is not necessary as it is already ∞.

Definition 2.12 Admission for Composition - Restricted Case

We first consider composition in the restricted case of length 1 layouts for the second layout.

Let $\mathbf{S} = (M_0, M_1, ..., M_\alpha)$ be a shape tuple, let $M = M_0 \cdot M_1 \cdot ... \cdot M_\alpha$, 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 $\hat{M} = r \cdot \hat{M}'$.
2. With respect to its induced factorization, $\hat{M}'$ is weakly left divisible by N.

Definition 2.13 Composition - Restricted Case

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

Suppose that $\mathbf{S} = (M_0, M_1, ..., M_\alpha)$ is a shape tuple, and $B = (N) : (r)$ is a layout of length 1 such that $\{\mathbf{S}, B\}$ is admissible for composition. Let $\mathbf{D} = (d_0, d_1, ..., d_\alpha)$ be any stride tuple and let $A = (\mathbf{S} : \mathbf{D})$ be a **coalesced** layout.

Note that in Jay Shah's original paper, the layout A was not specified to be coalesced. It will result in some compositions not being valid.

As in Definition 2.11, let $M = M_0 \cdot M_1 \cdot ... \cdot M_\alpha$ and $\hat{M} = r \cdot \hat{M}'$ with division index $0 \leq i \leq \alpha$. We separate the definition of $A \circ B$ into two cases.

First suppose $0 \leq i < \alpha$, so that

$$\begin{aligned} r &= M_0 \cdot M_1 \cdot ... \cdot M_{i-1} \cdot c \\ \hat{M}' &= \tfrac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{\alpha-1} \cdot \infty \end{aligned}$$

Then if $N \leq \frac{M_i}{c}$, we let $A \circ B = (N) : (cd_i)$.

Otherwise, there exists a $j \in [i+1, \alpha]$ such that $N = \frac{M_i}{c} \cdot ... \cdot M_{j-1} \cdot c'$, where $1 \leq c' < M_j$ if $j \neq \alpha$ (when $j = i+1$, $N = \frac{M_i}{c} \cdot c'$).

Note that here is an important fact that c' must be an integer because of the second condition for admission for composition, that is, $\hat{M}'$ is weakly left divisible by N. We must have $\frac{M_i}{c} \cdot ... \cdot M_{j-1}$ divides N, resulting in c' being an integer.

We let

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

If instead $i = \alpha$, then we have $r = M_0 \cdot M_1 \cdot \ldots \cdot M_{\alpha-1} \cdot c$ as before but $\hat{M}' = \infty$, and we let $A \circ B = (N) : (cd_\alpha)$.

Let' s look at this definition more closely.

Essentially, we are taking the one-dimensional coordinates $k \cdot r$ along the layout A where $k \in [0, N-1]$. Because we have $r = M_0 \cdot M_1 \cdot \ldots \cdot M_{i-1} \cdot c$, and c divides M_i.

Let first consider the case of $0 \leq i < \alpha$.

If $N \leq \frac{M_i}{c}$, then the first mode in the layout A is sufficient for dividing B. Consequently, the composition layout $A \circ B = (N) : (cd_i)$.

Otherwise if $N > \frac{M_i}{c}$, more modes in the layout A will be involved for dividing B, and consequently the composition layout

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

Let' s then consider the case of $i = \alpha$. We have $r = M_0 \cdot M_1 \cdot \ldots \cdot M_{\alpha-1} \cdot c$ and $\hat{M}' = \infty$. Here c is an positive integer that can be infinitely large. So only the last mode in the layout A is involved for dividing B, and consequently the composition layout $A \circ B = (N) : (cd_\alpha)$.

Note that by this definition, size$(A \circ B) =$ size(B). This is a critical property which we will use later.

Proposition 2.14

In the situation of Definition 2.13, we have that $f_{A \circ B} = \hat{f}_A \circ f_B$.

Proof

This more formally proves the intuition we explained to Definition 2.13.

We carry over the notation from Definition 2.13.

Given an 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 k-th position, where it is 1. Concretely,

$$\begin{aligned} \delta_0 &= \underbrace{(1, 0, 0, \ldots, 0)}_{\alpha+1} \\ \delta_1 &= \underbrace{(0, 1, 0, \ldots, 0)}_{\alpha+1} \\ &\vdots \\ \delta_\alpha &= \underbrace{(0, 0, 0, \ldots, 1)}_{\alpha+1} \end{aligned}$$

With respect to the the isomorphism of the extended layout A, we have

$$\hat{\imath} : \mathbb{N} \cong [0, M_0) \times [0, M_1) \times ... \times [0, M_{\alpha-1}) \times \mathbb{N}$$

Because $B = (N) : (r)$, we have

$$\begin{aligned} f_B(k) &= k \cdot r \\ &= M_0 \cdot M_1 \cdot ... \cdot M_{i-1} \cdot k \cdot c \end{aligned}$$

where $k \in [0, N-1]$.

Let first consider the case of $0 \leq i < \alpha$.

If $N \leq \frac{M_i}{c}$, i.e. $N \cdot c \leq M_i$, then we must have $k \cdot c < M_i$ for all $k \in [0, N-1]$. Because of the isomorphism of the extended layout A, we have

$$f_B(k) \mapsto \delta_i \cdot k \cdot c$$

Then we have

$$\begin{aligned} \left(\hat{f}_A \circ f_B\right)(k) &= \hat{f}_A\left(f_B(k)\right) \\ &= \hat{f}_A(\delta_i \cdot k \cdot c) \\ &= k \cdot c \cdot d_i \end{aligned}$$

According to Definition 2.13, we have

$$f_{A \circ B}(k) = k \cdot c \cdot d_i$$

Therefore, $f_{A \circ B} = \hat{f}_A \circ f_B$.

Otherwise if $N > \frac{M_i}{c}$, i.e. $N = \frac{M_i}{c} \cdot ... \cdot M_{j-1} \cdot c'$. Because of the isomorphism of the extended layout A, by definition, we have

$$\begin{aligned} f_B(k) &\mapsto \left(f_B(k) \mod M_0, \left\lfloor \tfrac{f_B(k)}{M_0} \right\rfloor \mod M_1, \left\lfloor \tfrac{f_B(k)}{M_0 \cdot M_1} \right\rfloor \mod M_2, ..., \left\lfloor \tfrac{f_B(k)}{M_0 \cdot M_1 \cdot ... \cdot M_{i-1}} \right\rfloor \mod M_i, ..., \left\lfloor \tfrac{f_B(k)}{M_0} \right. \right. \\ &= \left(0, 0, ..., (k \cdot c) \mod M_i, \left\lfloor \tfrac{k \cdot c}{M_i} \right\rfloor \mod M_{i+1}, ..., \left\lfloor \tfrac{k \cdot c}{M_i \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor \mod M_j, ..., \left\lfloor \tfrac{k \cdot c}{M_i \cdot M_{i+1} \cdot ... \cdot} \right. \right. \\ &= \left(0, 0, ..., \left(k \mod \tfrac{M_i}{c}\right) \cdot c, \left\lfloor \tfrac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}, ..., \left\lfloor \tfrac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor \mod M_j, ..., \left\lfloor \tfrac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ...} \right. \right. \end{aligned}$$

Note that here we used the property $(k \cdot c) \equiv \left(k \bmod \frac{M_i}{c}\right) \cdot c \pmod{M_i}$, if c divides M_i.

To see this, suppose $k \cdot c = p \cdot M_i + r$, where $0 \leq r < M_i$. Then we have $k = \frac{p \cdot M_i + r}{c} = p \cdot \frac{M_i}{c} + \frac{r}{c}$. Because c divides M_i, $\frac{r}{c}$ must be an integer. Thus, we have $(k \cdot c) \mod M_i = r$, and $\left(k \mod \frac{M_i}{c}\right) \cdot c = \frac{r}{c} \cdot c = r$. Therefore, $(k \cdot c) \mod M_i = \left(k \mod \frac{M_i}{c}\right) \cdot c$.

Further more, because $0 \leq k < N$, we have $0 \leq k \cdot c < N \cdot c = M_i \cdot ... \cdot M_{j-1} \cdot c'$, where $1 \leq c' < M_j$. Thus $0 \leq k \cdot c < M_i \cdot ... \cdot M_{j-1} \cdot M_j$.

When $c' > 1$, we have

$$\begin{aligned}\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_j} \right\rfloor &= \left\lfloor \frac{k\cdot c}{M_i\cdot M_{i+1}\cdot....\cdot M_j} \right\rfloor \\ &= 0\end{aligned}$$

$$\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_j} \right\rfloor \mod M_{j+1} = 0$$

and of course for any $l \in [j+1, \alpha-1]$, we have

$$\begin{aligned}\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_l} \right\rfloor &= \left\lfloor \frac{k\cdot c}{M_i\cdot M_{i+1}\cdot....\cdot M_l} \right\rfloor \\ &= 0\end{aligned}$$

$$\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_l} \right\rfloor \mod M_{l+1} = 0$$

Thus, we have

$$f_B(k) \mapsto \left(0,0,...,\left(k \mod \frac{M_i}{c}\right)\cdot c, \left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1},..., \left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_{j-1}} \right\rfloor \mod M_j, 0,0,...,0\right)$$

What' s more, because $k \le (N-1)$ and $\frac{M_i}{c}\cdot ...\cdot M_{j-1} = \frac{N}{c'}$, we have

$$\begin{aligned}\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_{j-1}} \right\rfloor &= \left\lfloor \frac{k}{\frac{N}{c'}} \right\rfloor \\ &= \left\lfloor \frac{k}{N}\cdot c' \right\rfloor \\ &\le \left\lfloor \frac{N-1}{N}\cdot c' \right\rfloor \\ &\le \lfloor c' \rfloor \\ &\le c'\end{aligned}$$

Because $c' < M_j$, we have

$$\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_{j-1}} \right\rfloor \mod M_j = \left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_{j-1}} \right\rfloor \mod c'$$

Thus, we have

$$f_B(k) \mapsto \left(0,0,...,\left(k \mod \frac{M_i}{c}\right)\cdot c, \left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1},..., \left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_{j-1}} \right\rfloor \mod c', 0,0,...,0\right)$$

Then we have

$$\begin{aligned}\left(\hat{f}_A \circ f_B\right)(k) &= \hat{f}_A\left(f_B(k)\right) \\ &= \hat{f}_A\left(0,0,...,\left(k \mod \frac{M_i}{c}\right)\cdot c, \left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1},..., \left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+1}\cdot....\cdot M_j}\right.\right. \\ &= 0\cdot d_0 + 0\cdot d_1 + ... + \left(k \mod \frac{M_i}{c}\right)\cdot c\cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right)\cdot d_{i+1} + ... + \left(\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{i+}}\right.\right. \\ &= \left(k \mod \frac{M_i}{c}\right)\cdot c\cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right)\cdot d_{i+1} + ... + \left(\left\lfloor \frac{k}{\frac{M_i}{c}\cdot M_{}}\right.\right.\end{aligned}$$

According to Definition 2.13, we have

$$f_{A\circ B}(k) \quad = \left(k \mod \frac{M_i}{c}\right) \cdot c \cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right) \cdot d_{i+1} + ... + \left(\left\lfloor \frac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor \mod c'\right) \cdot d_j$$

Therefore, $f_{A\circ B} = \hat{f}_A \circ f_B$.
When $c' = 1$, we have $0 \leq k \cdot c < N \cdot c = M_i \cdot ... \cdot M_{j-1} \cdot c' = M_i \cdot ... \cdot M_{j-1}$.
Thus, we have

$$\begin{aligned} \left\lfloor \frac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor &= \left\lfloor \frac{k \cdot c}{M_i \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor \\ &= 0 \end{aligned}$$

$$\left\lfloor \frac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-1}} \right\rfloor \mod M_j \quad = 0$$

Thus, we have

$$f_B(k) \quad \mapsto \left(0, 0, ..., \left(k \mod \frac{M_i}{c}\right) \cdot c, \left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}, ..., \left\lfloor \frac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-2}} \right\rfloor \mod M_{j-1}, 0, 0, ..., 0\right)$$

Then we have

$$\begin{aligned} \left(\hat{f}_A \circ f_B\right)(k) &= \hat{f}_A\left(f_B(k)\right) \\ &= \hat{f}_A\left(0, 0, ..., \left(k \mod \frac{M_i}{c}\right) \cdot c, \left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}, ..., \left\lfloor \right. \right. \\ &= 0 \cdot d_0 + 0 \cdot d_1 + ... + \left(k \mod \frac{M_i}{c}\right) \cdot c \cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right) \cdot d_{i+1} + ... + \left(\left\lfloor \right. \right. \\ &= \left(k \mod \frac{M_i}{c}\right) \cdot c \cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right) \cdot d_{i+1} + ... + \left(\left\lfloor \right. \right. \end{aligned}$$

According to Definition 2.13, we have

$$f_{A\circ B}(k) \quad = \left(k \mod \frac{M_i}{c}\right) \cdot c \cdot d_i + \left(\left\lfloor \frac{k}{\frac{M_i}{c}} \right\rfloor \mod M_{i+1}\right) \cdot d_{i+1} + ... + \left(\left\lfloor \frac{k}{\frac{M_i}{c} \cdot M_{i+1} \cdot ... \cdot M_{j-2}} \right\rfloor \mod M_{j-1}\right) \cdot d$$

Therefore, $f_{A\circ B} = \hat{f}_A \circ f_B$.
Let' s then consider the case of $i = \alpha$.

$$\begin{aligned} f_B(k) &= k \cdot r \\ &= k \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1} \cdot c \end{aligned}$$

where $k \in [0, N-1]$.
Because of the isomorphism of the extended layout A, we have

$$f_B(k) \;\mapsto \delta_\alpha \cdot k \cdot c$$

Then we have

$$\begin{aligned}\left(\hat{f}_A \circ f_B\right)(k) &= \hat{f}_A\left(f_B(k)\right) \\ &= \hat{f}_A(\delta_\alpha \cdot k \cdot c) \\ &= k \cdot c \cdot d_\alpha\end{aligned}$$

According to Definition 2.13, we have

$$f_{A\circ B}(k) \;= k \cdot c \cdot d_\alpha$$

Therefore, $f_{A\circ B} = \hat{f}_A \circ f_B$.

Taken together, we have $f_{A\circ B} = \hat{f}_A \circ f_B$ for all the cases in Definition 2.13.

This concludes the proof.

One might ask why the second condition for admission for composition is necessary. If we don't have it, c' can be fractional and we can still define $A \circ B$ to be

$$A \circ B = \begin{cases} \left(\frac{M_i}{c}, M_{i+1}, \ldots, M_{j-1}, \lceil c' \rceil\right) : \left(cd_i, d_{i+1}, \ldots, d_{j-1}, d_j\right) & \text{if } c' > 1 \\ \left(\frac{M_i}{c}, M_{i+1}, \ldots, M_{j-1}\right) : \left(cd_i, d_{i+1}, \ldots, d_{j-1}\right) & \text{if } c' = 1 \end{cases}$$

It's not too difficult to show that we still have $f_{A\circ B} = \hat{f}_A \circ f_B$ for the domain of f_B when the length of B is 1.

However, the critical property $\text{size}(A \circ B) = \text{size}(B)$ will not hold in this case. As we will see later, without having this property, $f_{A\circ B} = \hat{f}_A \circ f_B$ cannot be true when B is multi-modal, i.e., the length of B is greater than 1.

Definition 2.16 Interval of Definition

In the situation of Definition 2.12, where layout B is of length 1, 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 $M' = M_0 \cdot M_1 \cdot \ldots \cdot M_{\alpha-1}$ 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 Composition - General Case

Let $\mathbf{S} = (M_0, M_1, \ldots, M_\alpha)$ be a shape tuple, and let $B = (N_0, N_1, \ldots, N_\beta) : (r_0, r_1, \ldots, r_\beta)$ be a layout, let $B_k = (N_k) : (r_k)$ for $0 \le k \le \beta$. Then we say that the pair $\{\mathbf{S}, B\}$ is admissible for composition if:

1. For all $0 \le k \le \beta$, the pair $\{\mathbf{S}, B_k\}$ is admissible for composition in the sense of Definition 2.12.
2. The interval of definition for the pairs $\{\mathbf{S}, B_k\}_{0\le k\le\beta}$ are disjoint.

In this case, if $\mathbf{D} = (d_0, d_1, \ldots, d_\alpha)$ is a stride tuple and $A = \mathbf{S} : \mathbf{D}$, then we define the composition $A \circ B$ to be the concatenated layout

$$A \circ B := \left(A \circ B_0, A \circ B_1, ..., A \circ B_\beta\right)$$

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

Theorem 2.18 Composition - General Case

In the situation of Definition 2.17, we have that $f_{A\circ B} = \hat{f}_A \circ f_B$.

Proof

By Definition 2.13, size$(A \circ B_k)$ = size$(B_k) = N_k$ for all $0 \leq k \leq \beta$. We have the following isomorphism for both the layout $A \circ B$ or the layout B.

$$\iota : [0, N_0 \cdot N_1 \cdot ... \cdot N_\beta) \cong [0, N_0) \times [0, N_1) \times ... \times [0, N_\beta)$$

Given any $x \in [0, N_0 \cdot N_1 \cdot ... \cdot N_\beta)$, because of the isomorphism ι, we have

$$x \quad \mapsto \left(x_0, x_1, ..., x_\beta\right)$$

By Lemma 2.19, we have

$$\begin{aligned} \hat{f}_A \circ f_B(x) &= \hat{f}_A\left(f_B(x)\right) \\ &= \hat{f}_A\left(f_{B_0}(x_0) + f_{B_1}(x_1) + ... + f_{B_\beta}(x_\beta)\right) \end{aligned}$$

By Definition 2.17, Lemma 2.19, and Definition 2.13, we have

$$\begin{aligned} f_{A\circ B}(x) &= f_{A\circ B_0}(x_0) + f_{A\circ B_1}(x_1) + ... + f_{A\circ B_\beta}(x_\beta) \\ &= \hat{f}_A \circ f_{B_0}(x_0) + \hat{f}_A \circ f_{B_1}(x_1) + ... + \hat{f}_A \circ f_{B_\beta}(x_\beta) \\ &= \hat{f}_A\left(f_{B_0}(x_0)\right) + \hat{f}_A\left(f_{B_1}(x_1)\right) + ... + \hat{f}_A\left(f_{B_\beta}(x_\beta)\right) \end{aligned}$$

Normally, we don' t have $\hat{f}_A\left(x_{A,0} + x_{A,1} + ... + x_{A,\beta}\right) = \hat{f}_A\left(x_{A,0}\right) + \hat{f}_A\left(x_{A,1}\right) + ... + \hat{f}_A\left(x_{A,\beta}\right)$, because the layout function $\hat{f}_A$ is not linear. For example, suppose $A = (2,3) : (1,4)$, and we have $\hat{f}_A(1) = 1$ and $\hat{f}_A(3) = 5$. $\hat{f}_A(3) = \hat{f}_A(1+1+1) \neq \hat{f}_A(1) + \hat{f}_A(1) + \hat{f}_A(1)$.

However, there are some special cases where the above equation holds. For example, for simplicity, suppose $\beta = \alpha$, if we have

$$\begin{aligned} x_{A,0} &\in [0, M_0) \\ x_{A,1} &\in \{0, 1\cdot M_0, 2\cdot M_0, ..., \infty \cdot M_0\} \cap [0, M_1) \\ x_{A,2} &\in \{0, 1\cdot M_0 \cdot M_1, 2\cdot M_0 \cdot M_1, ..., \infty \cdot M_0 \cdot M_1\} \cap [0, M_2) \\ &\vdots \\ x_{A,\alpha} &\in \{0, 1\cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}, 2\cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}, ..., \infty \cdot M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}\} \cap [0, M_\alpha) \end{aligned}$$

By definition,

$$\hat{f}_A(x) = (x \mod M_0) \cdot d_0 + \left(\left\lfloor \frac{x}{M_0} \right\rfloor \mod M_1\right) \cdot d_1 + ... + \left(\left\lfloor \frac{x}{M_0 \cdot M_1 \cdot ... \cdot M_{\alpha-1}} \right\rfloor \mod M_\alpha\right) \cdot d_\alpha$$

So in our case, we have

$$\begin{aligned}\hat{f}_A\left(x_{A,0}+x_{A,1}+...+x_{A,\beta}\right) &= \left(\left(x_{A,0}+x_{A,1}+...+x_{A,\beta}\right) \mod M_0\right)\cdot d_0+\left(\left\lfloor\frac{x_{A,0}+x_{A,1}+...+x_{A,\beta}}{M_0}\right\rfloor \mod\right.\\ &= \left(x_{A,0} \mod M_0\right)\cdot d_0+\left(\left\lfloor\frac{x_{A,1}}{M_0}\right\rfloor \mod M_1\right)\cdot d_1+...+\\ &= \hat{f}_A\left(x_{A,0}\right)+\hat{f}_A\left(x_{A,1}\right)+...+\end{aligned}$$

The idea of having the second condition for admission for composition, i.e., the interval of definition for the pairs $\{\mathbf{S}, B_k\}_{0\leq k\leq\beta}$ are disjoint, are exactly the same.

Because $r_k = M_0\cdot M_1\cdot ...\cdot M_{i_k-1}\cdot c$, for $x_k\in[0,N_k)$, we have

$$\begin{aligned}f_{B_k}(x_k) &\in [0,1\cdot r,2\cdot r,...,(N_k-1)\cdot r]\\ &= [0,M_0\cdot M_1\cdot ...\cdot M_{i_k-1}\cdot c,2\cdot M_0\cdot M_1\cdot ...\cdot M_{i_k-1}\cdot c,...,(N_k-1)\cdot M_0\cdot M_1\cdot ...\cdot M_{i_k-1}\cdot c]\end{aligned}$$

Because of the isomorphism of the layout A, we have

$$f_{B_k}(x_k) \mapsto \left(x_{A,0},x_{A,1},...,x_{A,\alpha}\right)$$

where

$$x_{A,i} = \left\lfloor\frac{f_{B_k}(x_k)}{M_0\cdot M_1\cdot ...\cdot M_{i-1}}\right\rfloor \mod M_i$$

Then we must have some integer p_k and q_k, $p_k<q_k$, where $x_{A,i}=0$ for $i<p_k$ and $i>q_k$.

The second condition for admission for composition ensures that $[p_k,q_k]$ are disjoint for all $0\leq k\leq\beta$. Therefore, we can have the equation:

$$\hat{f}_A\left(x_{A,0}+x_{A,1}+...+x_{A,\beta}\right)=\hat{f}_A\left(x_{A,0}\right)+\hat{f}_A\left(x_{A,1}\right)+...+\hat{f}_A\left(x_{A,\beta}\right)$$

This concludes the proof.

Going back to the discussion why the second condition for admission for composition in the restricted case in Proposition 2.14 is necessary.

Because the critical property $\text{size}(A\circ B)=\text{size}(B)$ will not hold, we will have two completely different isomorphisms for the layout $A\circ B$ and the layout B, respectively.

Lemma 2.19 Concatenation of Layouts

Let $C=(C_0,C_1,...,C_\gamma)$ be a concatenated layout. Let

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

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

$$\begin{array}{ccc}[0,\text{size}(C)) & \underset{\cong}{\overset{\iota}{\to}} & [0,\text{size}(C_0))\times\cdots\times[0,\text{size}(C_\gamma))\\ f_C\downarrow & & \downarrow(f_{C_0},\cdots,f_{C_\gamma})\\ \mathbb{N} & \overset{+}{\leftarrow} & \mathbb{N}\times\cdots\times\mathbb{N}\end{array}$$

Proof

If $C_0, \ldots, C_\gamma$ are all length 1 layouts, then this is immediate from Definition 2.3.

Concretely, suppose $C_k = (M_k) : (d_k)$ for all $0 \leq k \leq \gamma$. The concatenated layout becomes

$$\begin{aligned} C &= (C_0, C_1, \ldots, C_\gamma) \\ &= (M_0 : d_0, M_1 : d_1, \ldots, M_\gamma : d_\gamma) \\ &= (M_0, M_1, \ldots, M_\gamma) : (d_0, d_1, \ldots, d_\gamma) \end{aligned}$$

Because of the isomorphism of the layout C, we have

$$x \mapsto \left(x_0, x_1, \ldots, x_\gamma\right)$$

Then by definition, the concatenated layout function is

$$f_C(x) = x_0 \cdot d_0 + x_1 \cdot d_1 + \ldots + x_\gamma \cdot d_\gamma$$

For each of the length 1 layouts C_k, by definition, the layout function is

$$f_{C_k}(x_k) = x_k \cdot d_k$$

Therefore, we have

$$f_C(x) = f_{C_0}(x_0) + f_{C_1}(x_1) + \ldots + f_{C_\gamma}(x_\gamma)$$

In the case where some of the layouts C_k are not length 1, we can apply the same argument to each of the sublayouts C_k, and the result follows by induction.

Concretely, suppose C_k are not length 1 and $C_k = (C_{k,0}, C_{k,1}, \ldots, C_{k,\gamma_k})$, where $C_{k,0}, \ldots, C_{k,\gamma_k}$ are length 1 layouts. Based on what we have proved above, we have

$$f_{C_k}(x_k) = f_{C_{k,0}}(x_{k,0}) + f_{C_{k,1}}(x_{k,1}) + \ldots + f_{C_{k,\gamma_k}}(x_{k,\gamma_k})$$

where

$$x_k \mapsto \left(x_{k,0}, x_{k,1}, \ldots, x_{k,\gamma_k}\right)$$

Suppose the layout C can be maximally decomposed into layouts of length 1.

$$\begin{aligned} C &= (C_0, C_1, \ldots, C_\gamma) \\ &= (C_{0,0}, C_{0,1}, \ldots, C_{0,\gamma_0}, C_{1,0}, C_{1,1}, \ldots, C_{1,\gamma_1}, \ldots, C_{\gamma,0}, C_{\gamma,1}, \ldots, C_{\gamma,\gamma_\gamma}) \end{aligned}$$

Then we have

$$\begin{aligned} f_C(x) &= f_{C_{0,0}}(x_{0,0}) + f_{C_{0,1}}(x_{0,1}) + \ldots + f_{C_{0,\gamma_0}}(x_{0,\gamma_0}) + f_{C_{1,0}}(x_{1,0}) + f_{C_{1,1}}(x_{1,1}) + \ldots + f_{C_{1,\gamma_1}}(x_{1,\gamma_1}) + \ldots \\ &= f_{C_0}(x_0) + f_{C_1}(x_1) + \ldots + f_{C_\gamma}(x_\gamma) \end{aligned}$$

where

$$x \;\mapsto (x_0, x_1, \dots, x_\gamma)$$

This concludes the proof.

Definition 2.21 CUTLASS Admission for Composition - Restricted Case

The CUTLASS admission for composition in the restricted case is more restrictive.

Let $\mathbf{S} = (M_0, M_1, \dots, M_\alpha)$ be a shape tuple, let $M = M_0 \cdot M_1 \cdot \dots \cdot M_\alpha$, 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 $\hat{M} = r \cdot \hat{M}'$.
2. With respect to its induced factorization, $\hat{M}'$ is left divisible by N.

Note that the second condition is the left divisibility, instead of the weak left divisibility in Definition 2.12.

For example, suppose $A = (8, 6, 8) : (1, 16, 108)$ and $B = (8) : (4)$. According to Definition 2.12, $A \circ B = (2, 4) : (4, 16)$. However, if we run composition for A and B in CUTLASS, we will encounter an error because CUTLASS requires left divisibility for the second condition.

More specifically, in the CUTLASS composition layout algebra implementation, we have

The reason why CUTLASS enforces this is because of the logical division operation. Without this restriction, the logical division operation will not be defined in some cases.

By-Mode Composition

In some cases, we would like to perform composition for each mode of the layout separately.

Let the layout A be a concatenation of layouts $A = (A_0, A_1, \dots, A_\alpha)$, and B be a tile of layouts $B = \langle B_0, B_1, \dots, B_\alpha \rangle$. Note that B, a tiler, is just a tuple of layouts instead of a concatenated layout. A tiler may consists of one or more than one layouts. Then we define the by-mode composition $A \circ B$ to be the layout

$$A \circ B := (A_0 \circ B_0, A_1 \circ B_1, \dots, A_\alpha \circ B_\alpha)$$

In some cases, we also have the layout A be a concatenation of layouts $A = (A_0, A_1, \dots, A_\alpha)$, and B be a tile of layouts $B = \langle B_0, B_1, \dots, B_\beta \rangle$, where $\beta \geq \alpha$. In this case, we define the by-mode composition $A \circ B$ to be the layout

$$A \circ B := (A_0 \circ B_0, A_1 \circ B_1, \dots, A_\alpha \circ B_\alpha)$$

Non-Integral Layout Composition

Similar to the non-integral layout complementation, the non-integral layout composition can be derived by coalescing the the layout so that the layout becomes integral. All the properties of the integral layout composition derived above can then be used.

Implication of Composition

The composition operation is usually used for selecting a sublayout from a layout. The sublayout can be strided even for each mode if the by-mode composition is performed.

Logical Division

Definition 2.22 Logical Division

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

$$A/B := A \circ (B, \text{complement}(B, M))$$

Note that here the conditions of admission for composition follows Definition 2.21 rather than Definition 2.12.

Implicitly Lemma 2.23 is used in Definition 2.22.

Lemma 2.23 Logical Division Implication

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

We denote $A = \mathbf{S} : \mathbf{D} = (M_0, M_1, \dots, M_\alpha) : (d_0, d_1, \dots, d_\alpha)$, and $B = (N_0, N_1, \dots, N_\beta) : (r_0, r_1, \dots, r_\beta)$. Let

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

be the automorphism such that $B^\varphi := (N_{\varphi(0)}, N_{\varphi(1)}, \dots, N_{\varphi(\beta)}) : (r_{\varphi(0)}, r_{\varphi(1)}, \dots, r_{\varphi(\beta)})$ is sorted.

Then by Definition 2.6, we have

$$\begin{aligned} B' &= \text{complement}(B, M) \\ &= \left(r_{\varphi(0)}, \tfrac{r_{\varphi(1)}}{N_{\varphi(0)} r_{\varphi(0)}}, \tfrac{r_{\varphi(2)}}{N_{\varphi(1)} r_{\varphi(1)}}, \dots, \tfrac{r_{\varphi(\beta)}}{N_{\varphi(\beta-1)} r_{\varphi(\beta-1)}}, \tfrac{M}{N_{\varphi(\beta)} r_{\varphi(\beta)}}\right) : (1, N_{\varphi(0)} r_{\varphi(0)}, N_{\varphi(1)} r_{\varphi(1)}, \dots, N_{\varphi(\beta-1)} r_{\varphi(\beta-1)}, \end{aligned}$$

Now we denote each mode of B' as

$$B'_k = \begin{cases} \left(r_{\varphi(0)}\right) : (1) & \text{if } k = 0 \\ \left(\tfrac{r_{\varphi(k)}}{N_{\varphi(k-1)} r_{\varphi(k-1)}}\right) : (N_{\varphi(k-1)} r_{\varphi(k-1)}) & \text{if } 1 \le k \le \beta \\ \left(\tfrac{M}{N_{\varphi(\beta)} r_{\varphi(\beta)}}\right) : (N_{\varphi(\beta)} r_{\varphi(\beta)}) & \text{if } k = \beta + 1 \end{cases}$$

for $k \in [0, \beta + 1]$.

Because the pair $\{\mathbf{S}, B\}$ is admissible for composition, for each mode in B, $B_k = (N_k) : (r_k)$ for $k \in [0, \beta]$, by Definition 2.17, the pair $\{\mathbf{S}, B_k\}$ is admissible for composition. Therefore, by Definition 2.12, M is left divisible by r_k and the quotient $\frac{M}{r_k}$ is left divisible (not weakly left divisible) by N_k for all $k \in [0, \beta]$.

It is trivial to see M is left divisible by 1. Let' s see if M is also left divisible by $N_{\varphi(k-1)} r_{\varphi(k-1)}$ for all $k \in [1, \beta + 1]$.

Suppose $\varphi(k-1) = h$ and Because M is left divisible by r_h, we have

$$\begin{aligned} r_{\varphi(k-1)} &= r_h \\ &= M_0 \cdot M_1 \cdot ... \cdot M_{i_{k-1}-1} \cdot c_{k-1} \end{aligned}$$

where c_k divides M_{i_k}.

$$\begin{aligned} N_{\varphi(k-1)} &= N_h \\ &= \frac{M_{i_k}}{c_{k-1}} \cdot M_{i_{k-1}+1} \cdot ... \cdot M_{j_{k-1}-1} \cdot c'_{k-1} \end{aligned}$$

where c'_{k-1} divides $M_{j_{k-1}}$.

Thus, M is also left divisible by $N_{\varphi(k-1)} r_{\varphi(k-1)}$, because

$$\begin{aligned} N_{\varphi(k-1)} r_{\varphi(k-1)} &= M_0 \cdot M_1 \cdot ... \cdot M_{i_{k-1}-1} \cdot c_{k-1} \cdot \frac{M_{i_k}}{c_{k-1}} \cdot M_{i_k+1} \cdot ... \cdot M_{j_{k-1}-1} \cdot c'_{k-1} \\ &= M_0 \cdot M_1 \cdot ... \cdot M_{j_{k-1}-1} \cdot c'_{k-1} \end{aligned}$$

where c'_{k-1} divides $M_{j_{k-1}}$.

Next, we will have to show M is left divisible by $\frac{r_{\varphi(k)}}{N_{\varphi(k-1)} r_{\varphi(k-1)}}$ for all $k \in [1, \beta]$.

$$r_{\varphi(k)} = M_0 \cdot M_1 \cdot ... \cdot M_{i_k-1} \cdot c_k$$

Because $r_{\varphi(k)} \geq N_{\varphi(k-1)} r_{\varphi(k-1)}$, we must have $i_k \geq j_{k-1}$. Thus,

$$\begin{aligned} \frac{r_{\varphi(k)}}{N_{\varphi(k-1)} r_{\varphi(k-1)}} &= \frac{M_0 \cdot M_1 \cdot ... \cdot M_{i_k-1} \cdot c_k}{M_0 \cdot M_1 \cdot ... \cdot M_{j_{k-1}-1} \cdot c'_{k-1}} \\ &= \frac{M_{j_{k-1}} \cdot M_{j_{k-1}+1} \cdot ... \cdot M_{i_k-1} \cdot c_k}{c'_{k-1}} \\ &= \frac{M_{j_{k-1}}}{c'_{k-1}} \cdot M_{j_{k-1}+1} \cdot ... \cdot M_{i_k-1} \cdot c_k \end{aligned}$$

Thus, M is left divisible by $\frac{r_{\varphi(k)}}{N_{\varphi(k-1)} r_{\varphi(k-1)}}$ for all $k \in [1, \beta]$.

It is trivial to see M is left divisible by $\frac{M}{N_{\varphi(\beta)} r_{\varphi(\beta)}}$.

Therefore, the pair $\{\mathbf{S}, B'_k\}$ is admissible for composition for all $k \in [0, \beta + 1]$.

By Definition 2.17, in order to show $\{\mathbf{S}, (B, \text{complement}(B, M))\}$ is admissible for composition, we also have to show the interval of definition for the pairs $\{\mathbf{S}, B_k\}_{0 \leq k \leq \beta}$ and $\{\mathbf{S}, B'_k\}_{0 \leq k \leq \beta+1}$ are disjoint.

By Proposition 2.7, the concatenated layout $(B, \text{complement}(B, M))$ is automatically satisfied with the disjoint argument.

Therefore, $\{\mathbf{S}, (B, \text{complement}(B, M))\}$ is admissible for composition.

This concludes the proof.

Note that in Definition 2.22, if the conditions of admission for composition follows Definition 2.12, our proof above will not be valid. That' s why CUTLASS enforces the conditions of admission for composition follows Definition 2.21.

By-Mode Logical Division

In some cases, we would like to perform logical division for each mode of the layout separately.

Let the layout A be a concatenation of layouts $A = (A_0, A_1, \ldots, A_\alpha)$, and B be a tile of layouts $B = \langle B_0, B_1, \ldots, B_\alpha \rangle$. Note that B is just a tuple of layouts instead of a concatenated layout. Then we define the by-mode logical division A/B to be the layout

$$\begin{aligned} A/B &:= (A_0/B_0, A_1/B_1, \ldots, A_\alpha/B_\alpha) \\ &= (A_0 \circ (B_0, \text{complement}(B_0, M)), A_1 \circ (B_1, \text{complement}(B_1, M)), \ldots, A_\alpha \circ (B_\alpha, \text{complemen} \\ &= ((A_0 \circ B_0, A_0 \circ \text{complement}(B_0, M)), (A_1 \circ B_1, A_1 \circ \text{complement}(B_1, M)), \ldots, (A_\alpha \circ B_\alpha, A_\alpha \circ \text{com} \end{aligned}$$

In some cases, we also have A be a concatenation of layouts $A = (A_0, A_1, \ldots, A_\alpha)$, and B be a tile of layouts $B = \langle B_0, B_1, \ldots, B_\beta \rangle$, where $\alpha \geq \beta$. In this case, we define the by-mode logical division A/B to be the layout

$$\begin{aligned} A/B &:= \left(A_0/B_0, A_1/B_1, \ldots, A_\beta/B_\beta, A_{\beta+1}, \ldots, A_\alpha\right) \\ &= \left(A_0 \circ (B_0, \text{complement}(B_0, M)), A_1 \circ (B_1, \text{complement}(B_1, M)), \ldots, A_\beta \circ \left(B_\beta, \text{complemen}\right.\right. \\ &= \left(\left(A_0 \circ B_0, A_0 \circ \text{complement}(B_0, M)\right), \left(A_1 \circ B_1, A_1 \circ \text{complement}(B_1, M)\right), \ldots, \left(A_\beta \circ B_\beta, A_\beta \circ \text{com}\right.\right. \end{aligned}$$

Logical Division Variants

In by-mode logical division, it is inconvenient to select a multi-dimensional tile. Given logical division A/B,

$$A/B = \left(\left(A_0 \circ B_0, A_0 \circ \text{complement}(B_0, M)\right), \left(A_1 \circ B_1, A_1 \circ \text{complement}(B_1, M)\right), \ldots, \left(A_\beta \circ B_\beta, A_\beta \circ \text{com}\right.\right.$$

We would like to iterate through each multi-dimensional sublayout $\left(A_0 \circ B_0, A_1 \circ B_1, \ldots, A_\beta \circ B_\beta\right)$ from A/B by indexing.

So zipped division is introduced by rearranging the layout A/B to

$$\begin{aligned} \text{zipped\_division}(A, B) &= \left(\left(A_0 \circ B_0, A_1 \circ B_1, \ldots, A_\beta \circ B_\beta\right), \left(A_0 \circ \text{complement}(B_0, M), A_1 \circ \text{complement}(B_1\right.\right. \\ &= \left(A \circ B, \left(A_0 \circ \text{complement}(B_0, M), A_1 \circ \text{complement}(B_1, M), \ldots, A_\beta\right.\right. \end{aligned}$$

In this way, we can iterate through each multi-dimensional sublayout $A \circ B = \left(A_0 \circ B_0, A_1 \circ B_1, \ldots, A_\beta \circ B_\beta\right)$ from zipped_division(A, B) by indexing on the second mode. Note because $A_0 \circ B_0$ has the same domain as B_0, $A_1 \circ B_1$ has the same domain as B_1, and so on, the domain or shape of $A \circ B$ becomes very predictable, which is $(\text{shape}(B_0), \text{shape}(B_1), \ldots, \text{shape}(B_\beta))$. The shape

of $\big(A_0 \circ \text{complement}(B_0, M), A_1 \circ \text{complement}(B_1, M), \ldots, A_\beta \circ \text{complement}(B_\beta, M)\big)$ is less predictable, but its 1D size is predictable, which is $\left(\frac{M}{\text{size}(B_0)}, \frac{M}{\text{size}(B_1)}, \ldots, \frac{M}{\text{size}(B_\beta)}\right)$.

CuTe implemented four logical division variants, including logical divide, zipped divide, tiled divide, and flat divide. Assuming the tiler is a tuple of two layouts, the logical division variants are defined as follows:

Implication of Logical Division

Because we have previously proved that $(B, \text{complement}(B, M))$ is a layout that is a bijection $[0, M) \cong [0, M)$, the logical division $A/B := A \circ (B, \text{complement}(B, M))$, where M is the size of A, is a layout that has the same domain as A and consequently also the same codomain as A. This means that the way the original layout A maps from the coordinates $[0, M)$ to the integers is scrambled or permutated to a new layout, i.e., the logical division A/B.

Because of the definition of composition,

$$\begin{aligned} A/B &:= A \circ (B, \text{complement}(B, M)) \\ &= (A \circ B, A \circ \text{complement}(B, M)) \end{aligned}$$

$A \circ B$ is the layout that selects a sublayout, i.e., a tile, from A based on the layout B, and $A \circ \text{complement}(B, M)$ that repeats the $A \circ B$ layout to fill the domain and codomain of A.

Logical division informs us how to extract a tile using the tiler B from a layout A, and how to repeat the tile to fill the domain and codomain of A, or how to select a tile from the consequent repeated layout using indexing.

Logical Product

Definition 4.1 Logical Product

Given a tiler and a layout, we could compute how the repeat layout that repeats the tiler to fill the domain and codomain of the layout using logical division. Similarly, given a tiler and a repeat layout, we could compute the resulting layout that is a tile of the repeat layout using logical product.

Let A and B be layouts, $M = \text{size}(A)\text{cosize}(B)$. Suppose that the pairs $\{A, M\}$ and $\{\text{complement}(A, M), B\}$ are admissible (for complementation and composition, respectively). Then we define the logical product $A \times B$ to be the layout

$$A \times B := (A, \text{complement}(A, M) \circ B)$$

The complementation of a layout A finds a complement layout $\text{complement}(A, M)$ with a positive integer M so that when the two layouts are concatenated, such as $(A, \text{complement}(A, M))$, the new layout is a bijection $[0, M) \cong [0, M)$. However, the codomain of the resulting layout we want to create might not need to be $[0, M)$. The layout B can have $\text{cosize}(B)$ that is much larger than $\text{size}(B)$. Consequently, after repeating the layout A using $\text{complement}(A, M) \circ B$, the resulting layout $A \times B$ will fill up to the codomain M but potentially strided. In addition, even if $\text{size}(B) = \text{cosize}(B)$, we could permute $\text{complement}(A, M)$ by using the composition operation, because there can be multiple layouts whose cosize is $\text{cosize}(B)$.

Because $\text{complement}(A, M) \circ B$ is a sublayout of $\text{complement}(A, M)$, the layout concatenation $(A, \text{complement}(A, M) \circ B)$ remains a valid layout.

By-Mode Logical Product

In some cases, we would like to perform logical division for each mode of the layout separately. For example, if we would like to tile a 2D layout with a 2D tiler, we would like to perform logical product for each mode of the layout separately.

Let the layout A be a concatenation of layouts $A = (A_0, A_1, \dots, A_\alpha)$, and B be a tile of layouts $B = \langle B_0, B_1, \dots, B_\alpha \rangle$. Note that B is just a tuple of layouts instead of a concatenated layout. Then we define the by-mode logical product $A \times B$ to be the layout

$$\begin{aligned} A \times B &:= (A_0 \times B_0, A_1 \times B_1, \dots, A_\alpha \times B_\alpha) \\ &= (A_0, \text{complement}(A_0, M) \circ B_0, A_1, \text{complement}(A_1, M) \circ B_1, \dots, A_\alpha, \text{complement}(A_\alpha, M) \circ B \\ &= ((A_0, \text{complement}(A_0, M) \circ B_0), (A_1, \text{complement}(A_1, M) \circ B_1), \dots, (A_\alpha, \text{complement}(A_\alpha, M) \circ \end{aligned}$$

Logical Product Variants

Similar to logical division, there are also variants of logical product, including logical product, zipped product, tiled product, and flat product, for the convenience of tile selection and iteration. Assuming the tiler is a tuple of two layouts, the logical product variants are defined as follows:

Implication of Logical Product

Similar to logical division, the logical product $A \times B$ informs the us what the layout is after repeating the original layout using the tiler.

Permutation Expressible As Layout Functions

This section explains how to retrieve all permutations that are expressible as layout functions in a structured way. This is important because some permutation algebra used in CUTLASS and CuTe, such as swizzle, does not seem to be expressed as layout function. The basic language of category theory is used to describe the process.

Definition 3.1 Ordered Factorization

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

For example, the set ob(**Fact**) includes expressions such as $[\,]$, $[2]$, $[3]$, $[22]$, $[23]$, $[32]$, $[232]$, etc.

Notation 3.3

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

Definition 3.4 Category of Ordered Factorizations

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

1. ob(**Fact**) is the set of objects of **Fact**.
2. For every expression $E = [p_1 \dots p_k]$ in ob(**Fact**) and every morphism of finite sets $\alpha : \underline{n} \to \underline{k}$, we have a morphism

$$E^\alpha = [p_{\alpha(1)} \dots p_{\alpha(n)}] \overset{\alpha_E}{\to} E = [p_1 \dots p_k]$$

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 expressions $E = [p_1 \dots p_k]$. Write

$$E^\alpha = [p_{\alpha(1)} \dots p_{\alpha(n)}] = [q_1 \dots q_n]$$

Let $\gamma = \alpha \circ \beta : \underline{m} \to \underline{k}$. Then the composition of morphisms

$$\alpha_E : E^\alpha = [p_{\alpha(1)} \dots p_{\alpha(n)}] \to E = [p_1 \dots p_k]$$

$$\beta_{E^\alpha} : E^\beta = [q_{\beta(1)} \dots q_{\beta(m)}] \to E^\alpha = [q_1 \dots q_n]$$

is given by $\gamma_E : E^\gamma \to E$, where we used that $[q_{\beta(1)} \dots q_{\beta(m)}] = [p_{\gamma(1)} \dots p_{\gamma(m)}]$.

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

To see why the composition of morphisms is associative, suppose we have morphisms of finite sets $\alpha : \underline{n} \to \underline{k}$, $\beta : \underline{m} \to \underline{n}$, and $\gamma : \underline{l} \to \underline{m}$. Then we have

$$\alpha \circ (\beta \circ \gamma) = (\alpha \circ \beta) \circ \gamma$$

To see why the composition of morphisms has identities, suppose for every n we have a morphism of finite sets $\mathrm{id}_{\underline{n}} : \underline{n} \to \underline{n}$, such that $\mathrm{id}_{\underline{n}}(i) = i$ for all $i \in \underline{n}$. Then we have

$$E^{\mathrm{id}_{\underline{n}}} = [p_{\mathrm{id}_{\underline{n}}(1)} \dots p_{\mathrm{id}_{\underline{n}}(n)}] \to E = [p_1 \dots p_n]$$

For every morphism $\alpha : \underline{n} \to \underline{k}$, we have

$$\alpha \circ \mathrm{id}_{\underline{n}} = \mathrm{id}_{\underline{k}} \circ \alpha = \alpha$$

Therefore, $\mathrm{id}_{\underline{n}}$ is the identity morphism for $\underline{n}$.

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.

In mathematics, a group is a set with an operation that associates an element of the set to every pair of elements of the set (as does every binary operation) and satisfies the following constraints: the operation is associative, it has an identity element, and every element of the set has an inverse element.

In this sense, the symmetric group is a set of all permutations of a set of k elements with an operation of composition of permutations (applying one permutation after another).

Suppose $E = [222]$. Then every permutation $\varphi \in \Sigma_3$ defines an automorphism $E^\varphi = E \to E$ in **Fact**.

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

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 **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 the identity morphism $[\mathrm{id}_{\underline{k}}]$.

Then for every expression $E = [p_1 \dots p_k]$ of length k, we have a functor

$$F_E : \textbf{FinSet}^{/\underline{k}} \to \textbf{Fact}$$

that sends the object $[\alpha : \underline{n} \to \underline{k}]$ to the expression 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.
Suppose we have an object of morphism $[\alpha : \underline{n} \to \underline{k}]$ and another object of morphism $[\beta : \underline{m} \to \underline{k}]$ in **FinSet**$^{/\underline{k}}$. Then the morphism of the overcategory is a commuting triangle, whose remaining morphism is $[\gamma : \underline{m} \to \underline{n}]$, that maps $[\alpha : \underline{n} \to \underline{k}]$ to $[\beta : \underline{m} \to \underline{k}]$.

The identity morphism $[\mathrm{id}_{\underline{k}}]$ is the final object of **FinSet**$^{/\underline{k}}$ because every other object of morphism in **FinSet**$^{/\underline{k}}$ has a unique morphism to $[\mathrm{id}_{\underline{k}}]$.

In this case, given an object of morphism $[\alpha : \underline{n} \to \underline{k}]$, the commuting triangle has a remaining morphism $[\gamma : \underline{k} \to \underline{n}]$, and we will have to show that this remaining morphism in the commuting triangle is unique.
Given $\underline{k} = \{1, 2, \dots, k\}$, the identity morphism maps $i \in \underline{k}$ to $i \in \underline{k}$. Given a morphism $[\alpha : \underline{n} \to \underline{k}]$, in which without loss of generality we assume $\alpha(i) = i$ for all $i \in [1, k]$, the remaining morphism $[\gamma : \underline{k} \to \underline{n}]$ must have $i = \alpha(i)$ for all $i \in [1, k]$, so that we have a commuting triangle that $i \in \underline{k} \overset{\gamma}{\to} \overset{\alpha}{\to} i \in \underline{k}$. Otherwise, for the remaining morphism $[\gamma : \underline{k} \to \underline{n}]$, if there exists an $i \in \underline{k}$ such that $i \neq \alpha(i)$, then the commuting triangle will not be valid because $i \in \underline{k} \overset{\gamma}{\to} \overset{\alpha}{\to} j \in \underline{k}$ and $i \neq j$. Therefore, the morphism of commuting triangle is unique.
By definition, let C and D be categories. A functor F from C to D is a mapping that

- associates each object X in C to an object $F(X)$ in D,

- associates each morphism $f : X \to Y$ in C to a morphism $F(f) : F(X) \to F(Y)$ in D such that
  - $F(\mathrm{id}_X) = \mathrm{id}_{F(X)}$ for every object X in C, and
  - $F(g \circ f) = F(g) \circ F(f)$ for all morphisms $f : X \to Y$ and $g : Y \to Z$ in C.

In the category $\mathbf{FinSet}^{/\underline{k}}$, we have $X = [\alpha : \underline{n} \to \underline{k}]$ and $Y = [\gamma : \underline{m} \to \underline{k}]$, $Z = [\mathrm{id}_{\underline{k}}]$, and the morphisms are commuting triangles $f = [\alpha] \to [\gamma]$, $g = [\gamma] \to [\mathrm{id}_{\underline{k}}]$, and $g \circ f = [\alpha] \to [\mathrm{id}_{\underline{k}}]$.

In the category **Fact**, by the functor F_E, we have $F_E(X) = E^\alpha = [p_{\alpha(1)} \dots p_{\alpha(n)}]$, $F_E(Y) = E^\gamma = [p_{\gamma(1)} \dots p_{\gamma(m)}]$, $F_E(Z) = E = [p_1 \dots p_k]$, and the morphisms are $F_E(f) = \alpha_E : E^\alpha \to E^\gamma$, $F_E(g) = \gamma_E : E^\gamma \to E$, and $F_E(g) \circ F_E(f) = \alpha_E : E^\alpha \to E$.

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 the infinite set of primes $\{2, 3, 5 \dots\}$, 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 $E = [p_1 \dots p_k]$ to the morphism $E_\bullet : k \to \mathcal{P}$ given by $i \mapsto p_i$.
Under this equivalence, the functor F_E of Remark 3.7 identifies with the functor

$$\mathbf{FinSet}^{/k} \simeq \left(\mathbf{FinSet}^{/\mathcal{P}}\right)^{/E_\bullet} \to \mathbf{FinSet}^{/\mathcal{P}}$$

that forgets the map to $E_\bullet$.
To understand this, let' s consider the following example.
Suppose we have an object $E = [232]$ from **Fact**. Then we have a morphism $E_\bullet : \underline{3} \to \mathcal{P}$ given by

$$\begin{aligned} i = 1 &\mapsto p_1 = 2 \\ i = 2 &\mapsto p_2 = 3 \\ i = 3 &\mapsto p_3 = 2 \end{aligned}$$

Every object of morphisms in $\mathbf{FinSet}^{/\mathcal{P}}$ can be mapped from an object from **Fact** and thus we have an equivalence of categories $\mathbf{Fact} \simeq \mathbf{FinSet}^{/\mathcal{P}}$.
Because of the functor F_E of Remark 3.7, with the equivalence above, we have

$$\mathbf{FinSet}^{/\underline{k}} \to \mathbf{FinSet}^{/\mathcal{P}}$$

Definition 3.9

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

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

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

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

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

Suppose $E = [23]$, $\varphi(1) = 2$, $\varphi(2) = 1$, $\varphi(3) = 2$. Then $L_{(E,\varphi)} = (3,2,3) : (2,1,2)$. This layout seems strange because its layout function is not an injection mapping. However, it is still a valid layout by Definition 2.2.

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

We could now see why p_i is prime number. It' s used for constructing the stride tuple of any kind for the layout, because any natural number can be uniquely factored into a product of prime numbers.

Remark 3.11

Let $E = [p_1 \dots p_k]$ and $\alpha : \underline{n} \to \underline{k}$. Let $N = p_1 \cdot p_2 \cdot \dots \cdot p_k$ and $N^\alpha = p_{\alpha(1)} \cdot p_{\alpha(2)} \cdot \dots \cdot p_{\alpha(n)}$. In what follows, consider the canonical isomorphisms

$$\begin{array}{ll} [0,N) & \cong [0,p_1) \times [0,p_2) \times \dots \times [0,p_k) \\ [0,N^\alpha) & \cong [0,p_{\alpha(1)}) \times [0,p_{\alpha(2)}) \times \dots \times [0,p_{\alpha(n)}) \end{array}$$

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

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

that sends the basis vector δ_i of one vector space to the basis vector $\beta_{\alpha(i)}$ of the other for $i \in [1,n]$.

In particular, if α is itself a bijection, then $f_{(E,\alpha)}$ restricts to an automorphism of $[0,N)$.

Proof

We noticed that $f_{(E,\alpha)} : [0,N^\alpha) \to [0,N) \subset \mathbb{N}$. The domain of $f_{(E,\alpha)}$ is $[0,N^\alpha)$ and it is obvious. The codomain of $f_{(E,\alpha)}$ is $[0,N) \subset \mathbb{N}$, however, is less obvious.

$$f_{(E,\alpha)} = x_{\alpha_{(1)}} d_1 + x_{\alpha_{(2)}} d_2 + \dots + x_{\alpha_{(n)}} d_n$$

where $x_{\alpha_{(i)}} \in [0,p_{\alpha(i)})$ and $d_i = \prod_{j<\alpha(i)} p_j$.

So we have

$$\begin{array}{ll} \max\left(f_{(E,\alpha)}\right) & = (p_{\alpha(1)}-1)d_1 + (p_{\alpha(2)}-1)d_2 + \dots + (p_{\alpha(n)}-1)d_n \\ & = (p_{\alpha(1)}-1)\prod_{j<\alpha(1)} p_j + (p_{\alpha(2)}-1)\prod_{j<\alpha(2)} p_j + \dots + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \end{array}$$

Without losing generality, we assume $p_{\alpha(1)} \le p_{\alpha(2)} \le \dots \le p_{\alpha(n)}$. Then we have

$$
\begin{aligned}
\max\left(f_{(E,\alpha)}\right) &= (p_{\alpha(1)}-1)\prod_{j<\alpha(1)} p_j + (p_{\alpha(2)}-1)\prod_{j<\alpha(2)} p_j + ... + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&\leq p_{\alpha(1)}\prod_{j<\alpha(1)} p_j + (p_{\alpha(2)}-1)\prod_{j<\alpha(2)} p_j + ... + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&= \prod_{j<\alpha(2)} p_j + (p_{\alpha(2)}-1)\prod_{j<\alpha(2)} p_j + ... + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&= p_{\alpha(2)}\prod_{j<\alpha(2)} p_j + ... + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&\leq \prod_{j<\alpha(3)} p_j + ... + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&\leq ... \\
&\leq p_{\alpha(n)}\prod_{j<\alpha(n-1)} p_j + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&= \prod_{j<\alpha(n)} p_j + (p_{\alpha(n)}-1)\prod_{j<\alpha(n)} p_j \\
&= p_{\alpha(n)}\prod_{j<\alpha(n)} p_j \\
&= \prod_{j\leq\alpha(n)} p_j \\
&\leq \prod_{j\leq k} p_j \\
&= N
\end{aligned}
$$

Thus, we have $f_{(E,\alpha)} : [0, N^\alpha) \to [0, N) \subset \mathbb{N}$.

Because $f_{(E,\alpha)}$ is a multilinear function, and because of the canonical isomorphisms, $f_{(E,\alpha)}$ can be described as the multilinear function

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

We denote a vector space $V = [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$ and a vector space $W = [0, p_1) \times [0, p_2) \times ... \times [0, p_k)$. Then the layout function $f_{(E,\alpha)}$ is a linear map $V \to W$.

Suppose $v_1, v_2, av_1, bv_2, av_1 + bv_2 \in V$, $f_{(E,\alpha)}(v_1) = w_1$, and $f_{(E,\alpha)}(v_2) = w_2$. Then we have

$$
\begin{aligned}
f_{(E,\alpha)}(v_1) &= v_{1,1}d_1 + v_{1,2}d_2 + ... + v_{1,n}d_n \\
f_{(E,\alpha)}(v_2) &= v_{2,1}d_1 + v_{2,2}d_2 + ... + v_{2,n}d_n \\
f_{(E,\alpha)}(av_1) &= av_{1,1}d_1 + av_{1,2}d_2 + ... + av_{1,n}d_n \\
&= af_{(E,\alpha)}(v_1) \\
f_{(E,\alpha)}(bv_2) &= bv_{2,1}d_1 + bv_{2,2}d_2 + ... + bv_{2,n}d_n \\
&= bf_{(E,\alpha)}(v_2) \\
f_{(E,\alpha)}(av_1 + bv_2) &= (av_1 + bv_2)_1 d_1 + (av_1 + bv_2)_2 d_2 + ... + (av_1 + bv_2)_n d_n \\
&= af_{(E,\alpha)}(v_1) + bf_{(E,\alpha)}(v_2)
\end{aligned}
$$

So $f_{(E,\alpha)} : V \to W$ is indeed a linear (multilinear) map.

Given an index $1 \leq i \leq \alpha$, let $\delta_i \in \mathbb{N}^{\times\alpha}$ denote the coordinate that is zero everywhere except in the k-th position, where it is 1. Note here the indexing is 1-based, instead of the similar one used in Proposition 2.14 which is 0-based. δ_i is the basis vector of the vector space V for $1 \leq i \leq \alpha$.

We send δ_i to $f_{(E,\alpha)}$ for $1 \leq i \leq \alpha$. Then we have

$$\begin{aligned} f_{(E,\alpha)}(\delta_i) &= d_i \\ &= \prod_{j<\alpha(i)} p_j \end{aligned}$$

Given the canonical isomorphism $[0, N) \cong [0, p_1) \times [0, p_2) \times ... \times [0, p_k)$, we have the multilinear function $g : W \to \mathbb{N}$

$$g(w) = w_1 + w_2 p_1 + w_3 p_1 p_2 + ... + w_k \prod_{j<k} p_j$$

Given an index $1 \leq i \leq k$, let $\beta_i \in \mathbb{N}^{\times k}$ denote the coordinate that is zero everywhere except in the k-th position, where it is 1. β_i is the basis vector of the vector space W for $1 \leq i \leq k$.

Thus, we have

$$\begin{aligned} f_{(E,\alpha)}(\delta_i) &= \prod_{j<\alpha(i)} p_j \\ &= g(\beta_{\alpha(i)}) \end{aligned}$$

This means the basis vector δ_i in the vector space V is sent to the basis vector $\beta_{\alpha(i)}$ in the vector space W by the multilinear function.

Suppose $v = c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha \in V$. Then we have

$$\begin{aligned} f_{(E,\alpha)}(v) &= f_{(E,\alpha)}(c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha) \\ &= (c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha)_1 d_1 + (c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha)_2 d_2 + ... + (c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha)_\alpha \\ &= c_1 d_1 + c_2 d_2 + ... + c_\alpha d_\alpha \\ &= c_1 f_{(E,\alpha)}(\delta_1) + c_2 f_{(E,\alpha)}(\delta_2) + ... + c_\alpha f_{(E,\alpha)}(\delta_\alpha) \\ &= c_1 g(\beta_{\alpha(1)}) + c_2 g(\beta_{\alpha(2)}) + ... + c_\alpha g(\beta_{\alpha(\alpha)}) \\ &= g(c_1\beta_{\alpha(1)} + c_2\beta_{\alpha(2)} + ... + c_\alpha\beta_{\alpha(\alpha)}) \end{aligned}$$

Therefore, we have set up the basis vector mapping for the multilinear function $f_{(E,\alpha)} : V \to W$. Given $v = c_1\delta_1 + c_2\delta_2 + ... + c_\alpha\delta_\alpha \in V$, it maps to $w = c_1\beta_{\alpha(1)} + c_2\beta_{\alpha(2)} + ... + c_\alpha\beta_{\alpha(\alpha)} \in W$.

This concludes the proof.

Lemma 3.12

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.

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

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

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

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

Proof

Let $N = p_1 \cdot p_2 \cdot ... \cdot p_k$, $N^\alpha = p_{\alpha(1)} \cdot p_{\alpha(2)} \cdot ... \cdot p_{\alpha(n)}$, and $N^\gamma = p_{\gamma(1)} \cdot p_{\gamma(2)} \cdot ... \cdot p_{\gamma(m)}$. We use the canonical isomorphisms

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

to write the domains and codomains of the layout functions in question.

More specifically, we have

$$\begin{array}{ll} f_{(E,\alpha)} : [0, N^\alpha) & \to [0, N) \\ f_{(E^\alpha,\beta)} : [0, N^\gamma) & \to [0, N^\alpha) \\ f_{(E,\gamma)} : [0, N^\gamma) & \to [0, N) \end{array}$$

We are trying to equate the multilinear function

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

with the composition of the two multilinear functions

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

We denote a vector space $V = [0, p_{\alpha(1)}) \times [0, p_{\alpha(2)}) \times ... \times [0, p_{\alpha(n)})$, a vector space $W = [0, p_1) \times [0, p_2) \times ... \times [0, p_k)$, and a vector space $U = [0, p_{\gamma(1)}) \times [0, p_{\gamma(2)}) \times ... \times [0, p_{\gamma(m)})$. The basis vectors of V, W, and U are δ_i, σ_j, and τ_l for $1 \leq i \leq n$, $1 \leq j \leq k$, and $1 \leq l \leq m$.

Based on the basis vector mapping by Remark 3.11, given $u = c_1\tau_1 + c_2\tau_2 + ... + c_m\tau_m \in U$, by $f_{(E^\alpha,\beta)}$, it maps to $v = c_1\delta_{\beta(1)} + c_2\delta_{\beta(2)} + ... + c_m\delta_{\beta(m)} \in V$. Then by $f_{(E,\alpha)}$, it maps to $w = c_1\sigma_{\alpha(\beta(1))} + c_2\sigma_{\alpha(\beta(2))} + ... + c_m\sigma_{\alpha(\beta(m))} \in W$.

Given $u = c_1\tau_1 + c_2\tau_2 + ... + c_m\tau_m \in U$, by $f_{(E,\gamma)}$, because $\gamma = \alpha \circ \beta$, $\gamma(i) = \alpha(\beta(i))$, it maps to $w' = c_1\sigma_{\gamma(1)} + c_2\sigma_{\gamma(2)} + ... + c_m\sigma_{\gamma(m)} = c_1\sigma_{\alpha(\beta(1))} + c_2\sigma_{\alpha(\beta(2))} + ... + c_m\sigma_{\alpha(\beta(m))} \in W$.

Because $w = w'$, we have $f_{(E,\gamma)} = f_{(E,\alpha)} \circ f_{(E^\alpha,\beta)}$.

This concludes the proof.

In Lemma 3.12, the per-mode condition of admissibility for composition (Definition 2.12) is satisfied. To see this, we have

$$\begin{array}{ll} E & = [p_1 p_2 ... p_k] \\ E^\alpha & = [p_{\alpha(1)} p_{\alpha(2)} ... p_{\alpha(n)}] \\ E^\gamma & = [p_{\gamma(1)} p_{\gamma(2)} ... p_{\gamma(m)}] \end{array}$$

$$L_{(E,\alpha)} = \left(p_{\alpha(1)}, p_{\alpha(2)}, ..., p_{\alpha(n)}\right) : (d_1, d_2, ..., d_n)$$

where $d_i = \prod_{j<\alpha(i)} p_j$.

$$L_{(E^\alpha,\beta)} = \left(p_{\alpha(\beta(1))}, p_{\alpha(\beta(2))}, \cdots, p_{\alpha(\beta(m))}\right) : (d'_1, d'_2, \ldots, d'_m)$$

where $d'_i = \prod_{j<\beta(i)} p_{\alpha(j)}$. CuTe Layout Algebra - Lei Mao's Log Book_files Because

$$\begin{aligned} M &= p_{\alpha(1)} \cdot p_{\alpha(2)} \cdot \ldots \cdot p_{\alpha(n)} \\ &= \left(\prod_{j<\beta(i)} p_{\alpha(j)}\right) \cdot p_{\alpha(\beta(i))} \cdot p_{\alpha(\beta(i)+1)} \cdot \ldots \cdot p_{\alpha(n)} \\ &= \left(\prod_{j<\beta(i)} p_{\alpha(j)}\right) \cdot M' \end{aligned}$$

Thus, M is left divisible by d'_i, M' is weakly left divisible and also left divisible by $p_{\alpha(\beta(i))}$, and the per-mode condition of admissibility for composition is satisfied.

The disjointness condition in Definition 2.17 is satisfied when $\beta : \underline{m} \to \underline{n}$ is an injective function and may be violated when it is not.

When $\beta : \underline{m} \to \underline{n}$ is injective, we have $m \leq n$, and $\beta(i) \neq \beta(j)$ for $i \neq j$. By Definition 2.16, for each mode $i \in [1, m]$, we have $N_i = p_{\alpha(\beta(i))}$, $I_i = [d'_i, d'_i(N_i - 1)]$. $M' = p_{\alpha(1)} \cdot p_{\alpha(2)} \cdot \ldots \cdot p_{\alpha(n-1)}$. So the interval of definition is $J_i = I_i \cap [1, M')$. Because $d'_i = \prod_{j<\beta(i)} p_{\alpha(j)} \geq 1$, $d'_i(N_i - 1) = \prod_{j<\beta(i)} p_{\alpha(j)} \cdot (p_{\alpha(\beta(i))} - 1) = \prod_{j\leq\beta(i)} p_{\alpha(j)} - \prod_{j<\beta(i)} p_{\alpha(j)} < M'$. Thus, $J_i = I_i = [\prod_{j<\beta(i)} p_{\alpha(j)}, \prod_{j\leq\beta(i)} p_{\alpha(j)} - \prod_{j<\beta(i)} p_{\alpha(j)}]$. Suppose we have a different mode k, $k \neq i$. Then $J_k = I_k = [\prod_{j<\beta(k)} p_{\alpha(j)}, \prod_{j\leq\beta(k)} p_{\alpha(j)} - \prod_{j<\beta(k)} p_{\alpha(j)}]$. Because $\beta(i) \neq \beta(k)$, without losing generality, we assume $\beta(i) < \beta(k)$. Then we have

$$\begin{aligned} \prod_{j<\beta(k)} p_{\alpha(j)} - \left(\prod_{j\leq\beta(i)} p_{\alpha(j)} - \prod_{j<\beta(i)} p_{\alpha(j)}\right) &= \prod_{j<\beta(k)} p_{\alpha(j)} - \prod_{j\leq\beta(i)} p_{\alpha(j)} + \prod_{j<\beta(i)} p_{\alpha(j)} \\ &> 0 \end{aligned}$$

Thus, $J_i \cap J_k = \emptyset$ for any $i \neq k$. The disjointness condition is satisfied.

When $\beta : \underline{m} \to \underline{n}$ is not injective, we don' t have $\beta(i) \neq \beta(j)$ for $i \neq j$. The disjointness condition may be violated.

So Lemma 3.12 actually proves Theorem 2.18 for layouts that has any arbitrary strides (not yet any arbitrary shapes) of the second layout that satisfies Definition 2.17.

Definition 3.14

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

Let R : **Fact** $\to$ **FinSet** be the functor defined as follows:

1. Let $E = [p_1 p_2 \ldots p_k]$ be an object of **Fact** and let $N = p_1 \cdot p_2 \cdot \ldots \cdot p_k$. Then $R(E) = [0, N)$.

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, $R :$ **Fact** $\to$ **FinSet** does indeed define a functor since it respects the composition of morphisms and identities as well.

We note that, as mentioned previously, R does not contain every possible function expressible as a layout function in its image. However, it does contain every automorphism of $[0, N) \xrightarrow{\cong} [0, N)$ 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 $f = f_L$. 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 $(p_1, p_2, \dots, p_k)$ where the p_i are all prime numbers and $N = p_1 \cdot p_2 \cdot \dots \cdot p_k$.

In order for f_L to be an automorphism of $[0, N)$, the sorted L, L^φ, must be of the form

$$L^\varphi = \left(p_{\varphi(1)}, p_{\varphi(2)}, \dots, p_{\varphi(k)}\right) : \left(1, p_{\varphi(1)}, p_{\varphi(1)}p_{\varphi(2)}, \dots, \prod_{1 \le i < k} p_{\varphi(i)}\right)$$

for some permutation $\varphi \in \Sigma_k$. This means that if we let $\psi = \varphi^{-1}$ be the inverse permutation, then

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

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

This concludes the proof.

Remark 3.16

One way to interpret Proposition 3.15 is that if we take the maximal subgroupoid **Fact**$^\simeq$ inside **Fact**, i.e., the subcategory of all invertible morphisms, then

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

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

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

be the full subgroupoid of those objects $[p_1 p_2 \dots p_k]$ with $N = p_1 \cdot p_2 \cdot \dots \cdot p_k$, then the Σ_N^L consists of those morphisms in the image of R on **Fact**$_N^\simeq$. However, we see that Σ_N^L is closed under the operation of taking the group inverse (the objects taking permutations to their inverses are also

in $\mathbf{Fact}_N^{\simeq}$). 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 the symmetric group Σ_k. This corresponds to $\mathbf{Fact}_{p^k}^{\simeq}$ being a groupoid with a single object $[pp \dots p]$, i.e., a group.