

An Introduction to Rank-polymorphic Programming in Remora (Draft)

Olin Shivers, Justin Slepak, Panagiotis Manolios
Northeastern University

December 31, 2019

Abstract

Remora is a higher-order, rank-polymorphic array-processing programming language, in the same general class of languages as APL and J. It is intended for writing programs to be executed on parallel hardware.

We provide an example-driven introduction to the language and its general computational model, originally developed by Iverson for APL. We begin with Dynamic Remora, a variant of the language with a dynamic type system (as in Scheme or Lisp), to introduce the fundamental computational mechanisms of the language, then shift to Explicitly Typed Remora, a variant of the language with a static, dependent type system that permits the shape of the arrays being computed to be captured at compile time.

This article can be considered an introduction to the general topic of the rank-polymorphic array-processing computational model, above and beyond the specific details of the Remora language. A reader generally interested in the topic of the computational model that serves as the foundation for this entire class of languages should find the tutorial informative.

We do not address the details of type *inference* in Remora, that is, the assignment of explicit types to programs written without such annotations; this is ongoing research [4].

Introduction

The rank-polymorphic programming model was developed by Kenneth Iverson, first for his array-programming language APL [2], and then later refined for its successor J [3]. We have subsequently designed a statically typed, higher-order functional language, Remora, based on the same computational model. The primary aim of this article is to introduce the reader, in a gradual, gentle way, to programming in Remora. We'll also discuss the pragmatics of Remora programming, in particular, how the parallel semantics of the language permits efficient code to be written for execution on parallel hardware. Finally, the design of Remora can serve as an introduction to the general model of computation that it embodies, which may make it easier to learn other languages in the class. A formal semantics for Remora, both its dynamic semantics and static type system, is presented in other work [5, 6].

Rank-polymorphic languages are known for not requiring explicit iteration or recursion constructs. Instead, the “iteration space” of a program is made real, or “reified,” in the shape of its aggregate data structures: when a function that processes an individual element of this space is applied to such a data structure, it is automatically lifted by a general polymorphic mechanism to apply across all of the elements of the aggregate.

In this tutorial, we'll look at the three core mechanisms that exist in Remora that work together to constitute its control story¹:

- Frame polymorphism
- Principal-frame cell replication
- Reranking

The interplay of these mechanisms permits sophisticated Remora (or APL, or J) programmers to write programs that are startlingly succinct.

Additionally, we will explore Remora's static type system, which permits the language to describe, at compile time, the dimensions of the arrays computed by Remora programs. We'll begin this tutorial by avoiding any mention of static types; once the dynamic mechanisms of the language are understood, we'll move on to the issue of how to capture the shapes of arrays with a static semantics, in ways that respect the above three mechanisms

¹Other languages, such as Python, Matlab, or R, have *ad hoc* mechanisms that permit programmers to do some of the same things, but without the same generality and design integrity of languages centrally based on Iverson's computational model.

Everything is an array

In rank-polymorphic languages such as Remora, *all values are arrays*. That is, every Remora expression evaluates to an array. An array is a collection of data arranged in a hyper-rectangle of some given dimensionality. Every array comes with its constituent elements, and a shape. Array elements come from a separate universe of *atoms*; typical atoms are numbers, characters, booleans and functions. Permitting arrays of functions means that Remora is a higher-order functional language.

For example, consider a matrix that has two rows and three columns of integers

$$\begin{bmatrix} 7 & 1 & 2 \\ 2 & 0 & 5 \end{bmatrix}$$

We say that this matrix has *rank* 2—that is, it has two dimensions or axes of indexing—and *shape* [2, 3]. The shape of an array is a sequence (or, equivalently, list or vector) giving its dimensions.

As another example, suppose we have collected rainfall data showing the monthly rainfall for twelve months of the year, across fifteen years of data collection, for all fifty states of the USA. We could collect this data as a numeric array RF of rank 3 and shape [50, 15, 12].

In principle, we could pull out the rainfall for April (month 3) of year 6 for the state of Georgia (state #9) by indexing into the array with the appropriate indices: RF [9, 6, 3]. But well-written programs in rank-polymorphic languages do not operate on individual elements of arrays; as we'll see, programs operate on entire arrays. So indexing is, in fact, something upon which we frown.²

The rank of a scalar array is 0 and its shape is the empty vector []. Note that:

- The rank of an array is also the length of its shape, which is maintained in the case of scalar values.
- Multiplying together the numbers in the shape of an array tells us how many atoms the array contains. For example, the shape of our rainfall-data array is [50, 15, 12], so the array contains $50 \times 15 \times 12 = 9000$ elements.

²It's not impossible to do: as we'll see in a later section (page 30), Remora does have an indexing operator that works in a data-parallel way. It's best to think of indexing in Remora as a fairly heavyweight communications operation that permits programmers to shuffle entire collections of data, not as a means of accessing array elements one-at-a-time.

As a boundary case, consider the scalar array whose only element is the number 17. The shape vector for a scalar array is the empty vector `[]`; multiplying all the elements of the empty vector together produces 1, which is, indeed, the number of elements contained by a scalar array.

In Remora, a language with a LISP-like s-expression syntax, the primitive notation for writing a literal array is the array form, that gives the shape of the array followed by its elements listed in row-major order. So our two example arrays, above, along with the scalar 17, could be written in Remora as the constant expressions

```
(array [2 3] 7 1 2 2 0 5) ; Our 2x3 example matrix
(array [50 15 12]          ; Rainfall data
      8 14 10 10 ...)    ; 9,000 elements here
(array [] 17)             ; The scalar value seventeen
```

Note that Remora’s basic s-expression syntax uses square brackets as well as parentheses; these are notationally distinct. Note, also, the LISP comment syntax: all text from a semicolon to the end of a line is ignored.

Array-producing expressions can be assembled into larger arrays with the frame form:

```
(frame [ $d_1$  ...]  $e_1$  ...)
```

The first subform of a frame expression is a shape or list of dimensions $[d_1 \dots d_n]$. This is followed by as many expressions as the product of the d_i ; these must all produce arrays of identical shape $[d'_1 \dots d'_m]$. Once these expressions have been evaluated, their result arrays are assembled together to produce a final array of rank $n + m$ and shape $[d_1 \dots d_n d'_1 \dots d'_m]$.

For example, the following code defines `v` to be a 3-element vector, and `m` to be a two-row, three-column matrix whose two rows are each identical to `v`:

```
(define v (array [3] 8 1 7)) ; Shape [3]
(define m (frame [2] v v))   ; Shape [2 3]
```

Note the distinctions between the array and frame forms. The array form is for writing down array *constants*, that is, *literal* arrays; its subforms are literal atoms. The frame form causes run-time computation to occur: we evaluate the expressions that are its subforms to produce arrays that are then “plugged into” position in the given frame to make a larger, result array.

Now that we’ve introduced the array and frame forms, we’ll hide them from view at every turn by means of some convenient syntactic sugar:

- First, whenever an atom (that is, an array *element*) literal a appears in a syntactic context where we expect an expression³, it is taken to be a scalar array—that is, it is treated as shorthand for `(array [] a)`.
- Second, whenever a sequence of expressions occurs surrounded by square brackets in an expression context, it is treated as a frame form for a vector frame. That is, the expression `[e1 ... en]` is treated as shorthand for `(frame [n] e1 ... en)`.
- Finally, a frame whose component expressions are all array literals is, itself, collapsed to a single array term.

Thus we could write the scalar array 17 as expression 17, and the vector of the first five primes as expression `[2 3 5 7 11]`; our original example array could be written as

```
[[7 1 2]   ; A 2x3 matrix
 [2 0 5]]
```

This is *exactly* equivalent to the array-literal expression

```
(array [2 3] 7 1 2 2 0 5)
```

Likewise, we could write the truth table for $i \text{ xor } j \text{ xor } k$, using 0 for false and 1 for true, as the rank-3 array

```
;;; A 2x2x2 array
[[[0 1]   ; i=0 plane / j=0 row
 [1 0]]   ; i=0 plane / j=1 row

 [[1 0]   ; i=1 plane / j=0 row
 [0 1]]] ; i=1 plane / j=1 row
```

When using the square-bracket notation, the shape of the array is determined from the nesting structure of the expression. It's not allowed for two brother elements in a square-bracket array expression to have different shapes; they must match. Thus, the following “ragged” matrix is not a legal expression, as it doesn't have a well-defined shape:

³Remember: all expressions produce arrays.

```
[[7 1 2]
 [9 5]   ; Illegal -- row too short!
 [2 0 5]]
```

As we'll see later (page 43), there is a mechanism in Remora called a "box," that permits programmers to make ragged arrays, but we'll ignore this for now.

Functions operate on "cells" of input

In Remora, every function is defined to operate on arguments of a given rank and produce a result of a given rank; these are called the *cells* of the function application. For example, the addition operator `+` operates on two arguments, each of which is a scalar, that is, of rank 0.

```
(+ 3 4)
7
(+ 2 8)
10
```

In this example, and the examples to come, we'll show code and the result expressions it produces, in an "interactive" style, as if we were presenting Remora expressions and definitions to an interpreter: the input Remora expression will be indented, and the value produced will be displayed, flush left, on the following line.

As further examples, we could have a dot-product function `dot-prod` that operates on two arguments of rank 1; or a polynomial evaluation function `poly-eval` that operates on a vector (rank 1) giving the coefficients of a polynomial, and a scalar (rank 0) giving the x value where we are evaluating the polynomial:

```
(dot-product [2 0 1] [1 2 3])
5
;; Evaluate 2 + 0x - 3x^2 at x=1
(poly-eval [2 0 -3] 1)
-1
```

The argument ranks of a function are part of its static definition; when we define our own functions, we must specify them. We do this by tagging each parameter to the function with its rank. So, both x and y inputs to the `diff-square` function below are specified as being of rank 0:

```
(define (diff-square [x 0] [y 0])
  (- (* x x)
     (* y y)))

(diff-square 5 3)
```

16

Functions distribute over a frame of cells

The fundamental, core iteration mechanism of Remora is that any function defined to take arrays of rank r is automatically lifted by the language so that it can be applied to arrays of any rank $r' \geq r$. This is the polymorphism of “rank polymorphism.”

This lifting is accomplished by viewing an array as a *frame of cells*. Consider our 2×3 matrix:

$$\begin{bmatrix} 7 & 1 & 2 \\ 2 & 0 & 5 \end{bmatrix}$$

We can view this array three ways:

1. as a scalar frame containing a single 2×3 matrix cell;
2. as a vector frame containing two cells that are 3-vectors; or
3. as a 2×3 matrix frame containing six scalar cells which are the individual elements of the array.

In general, an array with rank r and shape $s = [d_1, \dots, d_r]$ can be viewed as a frame of cells in $r + 1$ ways, depending on where one splits the shape into the frame prefix and the cell suffix.

When we apply a function that consumes argument cells of rank r to an actual argument array of rank $r' \geq r$, we divide the input into a frame of cells: the cells have rank r , and the frame has rank $r' - r$. The shape of each cell is given by the last r dimensions of the array’s shape; the shape of the frame is the remaining, initial $r' - r$ dimensions of the array. That is, the function imposes a view on the argument array: we view the rank r' array as a collection of rank r argument cells.

The function is then applied, in parallel, to each argument cell; the results of all these independent applications (which must all have the same shape) are then collected into the frame to produce the final result.

For example, suppose we have a function `vmag` that takes a vector (that is, an array of rank 1) and returns its Euclidean length or magnitude:

```
(define (vmag [v 1]) ...)
(vmag [3 4])
5
(vmag [1 2 2])
3
```

Note that the `v` parameter to `vmag` is defined to take arguments of rank 1. If we apply `vmag` to a matrix, it is applied independently to each row of the matrix. That is, `vmag` views the matrix as a vector frame of vector cells. All the scalar results of these `vmag` applications are collected into the original argument's vector frame:

```
(vmag [[1 2 2]
      [2 3 6]])
[3 7]
```

Likewise, if we applied `vmag` to a six-dimensional array, it would be treated as a five-dimensional frame of vector cells; each such cell would have its length computed, and we would collect these scalar answers into the frame to produce a rank-5 array result.

Multiple arguments and frame agreement

The frame-distribution mechanism of function application applies just as well when a function has multiple arguments. For example, consider our polynomial-evaluation function, `poly-eval`, that takes a vector of coefficients and a scalar value at which we wish to evaluate the polynomial. Suppose we apply this function to a matrix and a vector

```
(poly-eval [[2 0 -3] ; two polynomials
           [5 -1 1]]
          [-2 1]) ; two x values
[-10 5]
```

The coefficient matrix has shape `[2, 3]`, and the vector of x values has shape `[2]`. Since `poly-eval` operates on vectors for its first argument, it views the matrix as a vector frame (shape `[2]`) of vector cells (shape `[3]`); likewise, it views its vector

Argument	Argument shape	Frame shape	Cell shape
#1 (coefficients)	[2, 3]	[2]	[3]
#2 (x)	[2]	[2]	[]

Table 1: The `poly-eval` function imposes a frame/cell view on each of its arguments. After removing the cell-shape suffix from the shape of each argument array, the remaining shape prefix is the argument’s frame. Both arguments in our example have the same frame, which means we can lift the function application across the two matching frames; this frame then provides the outer structure into which we collect the results of the distinct invocations of `poly-eval`.

of x values as a vector frame (shape [2]) of scalar cells (shape []). Note that once we’ve pulled off the cell-shape suffixes from the shapes of each argument, we are left with identical frame shapes: [2]. This is called “frame agreement,” which means we have a consistent frame across which to distribute the individual function applications. Thus, we evaluate the polynomial $2 + 0x - 3x^2$ at $x = -2$, and $5 - x + x^2$ at $x = 1$, collecting the results into the vector frame and producing final answer [-10 5]. Table 1 shows how the cell requirements of `poly-eval` guide the distribution of the function call over the collection of inputs it is given.

Remora’s lifting mechanism essentially distributes the `poly-eval` function across the two matching frames of arguments in a pointwise fashion, producing one function invocation for every cell of the frame. It is as if we had written

```
[ (poly-eval [2 0 -3] -2)
  (poly-eval [5 -1 1] 1) ]
[-10 5]
```

The principal frame and cell replication

Remora’s frame-based distribution mechanism is more general than simply requiring the frames of all argument arrays to match. The full rule is driven by the notion of an application’s “principal frame.” In a given function application, the argument frame with the *longest shape* is considered the principal frame; for the function application to be well-formed, the frame shapes of all other arguments must be a *prefix* of the principal frame’s. When distributing an argument’s cells across the cell-wise invocations of the function, if an argument’s frame has a shape

shorter than (a proper prefix of) the principal frame's, then the array is replicated into the missing dimensions to provide enough cells for the full frame of function applications.

For example, suppose we add a vector of 2 numbers to a 2×3 matrix:

```
(+ [10 20] [[8 1 3]
            [5 0 9]])
[[18 11 13]
 [25 20 29]]
```

The addition operator adds two scalars to produce a scalar result. Since the two frames have shapes $[2]$ and $[2, 3]$, the principal frame's shape is $[2, 3]$. The first argument's frame gets replicated from shape $[2]$ to shape $[2, 3]$. The way to think of this replication is that when we select a cell from this argument, for frame element $[i, j]$, we simply *drop* any suffix of the index not needed to index into the argument's actual frame—in this case, the column index j . This means that we match every column of the right argument's first row with the first element of the left vector: $10 + 8$, $10 + 1$, and $10 + 3$; likewise, we match the items of the right argument's second row with the second element of the left vector: $20 + 5$, $20 + 0$, $20 + 9$. Thus, we get one function application for each element of the principal frame, where the results are collected, producing an answer which is a 2×3 frame of scalar cells.

In short, the frame-agreement rule of Remora means that when we add a vector to a matrix, we add the first element of the vector to the first row of the matrix, the second element of the vector to the second row of the matrix, and so forth. (What if we want to add the first element of the vector to the first *column* of the matrix, and so forth? We'll come to this later.)

Likewise, if we add a matrix M to a three-dimensional array A , then we add element $M[i, j]$ to each element of plane i , row j of A ; that is, we add each scalar cell $M[i, j]$ to each scalar cell $A[i, j, k]$.

Given this rule, adding a scalar s to any array simply adds s to each element of the array:

```
(+ 10 [7 1 4])
[17 11 14]
(+ [7 1 4] 10)
[17 11 14]
```

If we wish to evaluate a collection of polynomials at a single x , we simply apply the `poly-eval` function to the collection and the x value. Whereas, if we

wish to evaluate a single polynomial at a collection of x values, we apply the function to the polynomial and the collection of x values:

```
;;; Evaluate two polynomials at the same x.
(poly-eval [[2 0 -3]          ; 2 + 0x - 3x^2
            [5 -1 1]]        ; 5 - x + x^2
            -1)              ; x = -1
[-1 7]
;;; Evaluate 2 + 0x - 3x^2 at four values of x.
(poly-eval [2 0 -3] [[0 1]
                    [2 3]])
[[ 2 -1]
 [-10 -25]]
```

The second example above shows that, in Remora, the shape of the output collection is determined by the shape of the input collection: When we pump a *matrix* of inputs through the same polynomial, we get a *matrix* of outputs, since the principal frame comes from the matrix argument.

In none of these cases do we need to write a loop, or index into a collection of data; this is managed for us by the rank-polymorphic lifting of the `poly-eval` function across its arguments.

It's worth noting that this replication mechanism only replicates *entire cells*. It does not replicate arrays that are smaller than one entire cell's worth of data up into a complete argument cell. Thus, it will not, for example, permit us to apply our vector dot-product function to a scalar, such as 0. In short, an actual argument must always provide enough data to a function to comprise at least one complete argument.

Frame-replication even applies to the function position

Because Remora is a higher-order functional language, we can write a general expression in the function position of a function-application expression; because expressions in Remora evaluate to arrays, this means that the function position of an application can be an *array* of functions. For example, `+` is a variable whose value in the top-level environment is a *scalar array* whose single element is the addition function—as described earlier, functions in Remora are atoms, that is, array elements, just as numbers, booleans, and characters are.

Thus, when we evaluate the expression that is the function position of a function application, we get an array (of functions), and *this array participates in the determination of the principal frame for the application, and is subject to frame replication just as the argument arrays are*. The ability to apply an array of different functions to an argument gives Remora a MIMD-style capability to its parallel semantics.

The function position takes scalar cells, which means that in the common case, when we apply a scalar function array (such as the + array) to a pair of argument arrays, it is replicated across all the applications.

But we can use non-scalar arrays of functions, as well. Here, we apply a matrix of functions to the single value 9, collecting the results into the matrix principal frame:

```
(define m [[square square-root] ; M is a 2x2 array
           [add1 sub1]]) ; of functions

(m 9) ; Apply all the functions to nine.
[[81 3]
 [10 8]]
```

Some functions take the entire argument as cell

The frame-of-cells story in Remora has a useful corner case: it is possible to specify that a particular parameter to a function takes its entire argument as a single cell. For example, the append function takes two arrays and appends them along their initial dimension. Appending two matrices appends the rows of the second matrix after the rows of the first matrix. So appending a 3×5 array and a 7×5 array produces a 10×5 result. Likewise, appending two three-dimensional arrays appends the planes of the second array after the planes of the first array: appending a $3 \times 2 \times 5$ array and a $4 \times 2 \times 5$ array produces a $7 \times 2 \times 5$ result.

```
(define m1 [[0 1]
            [2 3]])

(define m2 [[10 20]
            [30 40]])
```

```
;;; Append two 2x2 matrices; result is 4x2.  
(append m1 m2)
```

```
[[0 1]  
 [2 3]  
 [10 20]  
 [30 40]]
```

The `append` function is defined so that it consumes both of its arguments, of any rank, as a single cell; thus its frame is a scalar. When we define our own functions, we declare this by tagging a parameter with the special keyword `all` instead of a natural number for its cell rank:

```
(define (append [a all] [b all]) ...)
```

One way to view such a parameter is that we fix, not the *cell* rank of the parameter, but its *frame* rank: such a parameter has a scalar frame.

What if we want to append along a different axis of an array? For example, instead of appending the two previous matrices one above the other, suppose we wanted to append them side-by-side, producing:

```
[[0 1 10 20]  
 [2 3 30 40]]
```

We'll see how to do this in a following section.

Iteration with reduce, scan, fold and trace

An important operator with scalar frame rank is the higher-order `reduce` function, which maps an associative binary function `+` over the initial dimension of an array. That is, `reduce` takes an array of non-zero rank r and shape $[d_1, \dots, d_r]$, and considers it a collection of d_1 arrays of rank $r - 1$ and shape $[d_2, \dots, d_r]$. This collection of arrays $[a_1, \dots, a_{d_1}]$ is combined together using some provided function `+`, producing the result

```
(+ a1 (+ a2 (+ ... ad1)))
```

Since `reduce` requires the combining function `+` to be associative, we can think of the function as operating on all the items at once

`(+ a1 a2 ... ad1)`

For example, if we wish to sum the elements of a vector, we reduce it with `+`:

`(reduce + [1 4 9 16]) ; Sum of first four squares`
30

Similarly, if we wish to multiply the elements together, we reduce the same vector with `*`:

`(reduce * [1 4 9 16]) ; Product of first four squares`
576

The definition of `reduce` does not specify in what order these numbers will be multiplied. It may be done left-to-right

`(* (* (* 1 4) 9) 16)`

right-to-left

`(* 1 (* 4 (* 9 16)))`

or in any other order permitted by the axiom of associativity, such as

`(* (* 1 4) (* 9 16))`

If we reduce a matrix with `+`, we will add the first row to the second row, the third row, and so forth. So, in effect, we will sum each column:

`(reduce + [[1 2 3]
 [10 20 30]
 [100 200 300]])`
[111 222 333]

What if we want to sum along a different axis of the array? We'll see how to do this in a following section.

The associativity requirement for the combining operator restricts its type to $\alpha \times \alpha \rightarrow \alpha$. Note that the operator gets automatically lifted to operate on the subarrays if it is defined to take cells of smaller rank. Thus, our `+` operator, which fundamentally operates on scalars, was lifted to operate on vectors when it was used in the matrix example above.

The `reduce` operator must be applied to an array whose initial (or "iteration") dimension is non-zero. If we want to sum a matrix `m` that might have zero rows,

reduce has no data to feed the + operator, hence no way to produce a result. If this happens, Remora will report an error.⁴ In such a case, the programmer can use an alternate “reduce with zero” function, and explicitly provide a zero value, writing (reduce/zero + 0 m).

Remora also provides a fold function that uses a more general folding operator of type $\alpha \times \beta \rightarrow \beta$, combining items of type α into a running accumulator of type β . For example, we can compute the sum of the magnitudes of a collection of vectors with

```
(fold (λ ([v 1] [sum 0]) (+ sum (vmag v)))
      0
      [[1 2 2]      ; length 3
       [2 3 6]])   ; length 7
10
```

The advantage of using the less general reduce is that it permits the reduction to be performed in a parallel fashion; fold is serial. So we would be better off expressing the above calculation as:

```
(reduce + (vmag [[1 2 2]
                 [2 3 6]]))
```

The individual vmag computations can be executed in parallel, as can the additions of the final summation. This is unimportant in this example, where our matrix represents a collection of only two small vectors, but would be significant if we had either a large number of vectors, or our vectors were extremely long—that is, if our matrix had a large number of rows or columns (or both).

There is also an iscan operator that produces the “prefix sums” of an operator applied across the initial dimension of an array:

```
(iscan + [2 10 5]) ; Produce [2, 2+10, 2+10+5]
[2 12 17]

(iscan + [[1    2    3]
          [10   20   30]
          [100  200  300]])
[[1    2    3] ; row 1
 [11   22   33] ; row 1 + row 2
 [111  222  333] ; row 1 + row 2 + row 3
```

⁴A run-time error, in the case of Dynamic Remora; a compile-time type error, in the case of Typed Remora.

The `iscan` function provides an “interior scan:” the i th element of the result includes the i th element of the input collection.

An alternative class of scan functions are the “external” scans, which do *not* include the i th element of the input collection in element i of the result:

```
(scan/zero + 0 [2 10 5]) ; [0, 2, 2+10, 2+10+5]
[0 2 12 17]
```

Note that the exterior scan’s result is one element longer than its input: this is because an external scan includes the initial “zero” value in the result collection.

Remora actually provides three different reduce functions, eight scans, two folds and two traces. The trace functions are the “prefix-sum” version of the serial folds; that is, the trace functions have the same relationship to folding that the scan functions have to reductions. The small details differentiating these fifteen functions is something we won’t explore fully, here. In this tutorial, we’ll simply cover these differences as we introduce needed variants in our example code.

Remora’s reduce / scan / fold / trace set of operators provide an important component of its control story. The frame/cell lifting mechanism of the language enforces a separation of computation when we apply a low-rank operator to a high-rank collection of data. For example, when we apply the vector-magnitude operator `vmag` to a three-dimensional array of shape $7 \times 5 \times 3$, the 35 different applications of `vmag` all run independently of one another. This is desirable, as it permits all the different invocations of `vmag` to be executed in parallel. Sometimes, however, we need to perform a computation on a collection of data that somehow *combines together* the elements of the collection (a computation with what the scientific-computing community would call a “loop-carried dependency” when expressed in programming languages that have explicit loops). In these cases, we use `reduce`, `scan` or one of their serial brethren—that is their *raison d’être*.

Remora is a map/reduce architecture: an aside on parallelism

The use of `reduce` and `scan` characterises a big distinction between programming in a rank-polymorphic language like Remora and programming in a serial-array language like FORTRAN. In FORTRAN, we write loops and then hope the compiler can sort out which computations inside the loop are independent of the iteration order and can therefore be parallelised, and which computations have loop-carried dependencies, and so must be left serialised. In Remora, the notationally simple way to operate on a collection of data is simply to apply to the *collection* the

function that processes a *single* item: (`f collection`), and this default case is the parallel case. The *actual semantics* is parallel—we are not just emulating a serial semantics with a parallel implementation—so the compiler is licensed to perform all the per-item calculations in parallel; no heroic analyses are needed to divine this fact. On the other hand, the compiler has no difficulty spotting loop-carried dependencies when they do arise, because the programmer *explicitly marks* them by writing down one of the reduce / scan / fold / trace operators.

What is particularly nice about providing reductions and scans as a central iteration construct is that these operations are a parallelism slam-dunk. Suppose we are reducing a collection of n items on a parallel machine that has p processors. If we have more processors than data ($n < p$), we can do the operation in $O(\lg n)$ time. Since n measures a physical resource, then here in the physical universe, where most of us reside, it is entirely reasonable to think of this $O(\lg n)$ time simply as being able to do a reduction in constant time, for some constant less than 100. Alternately, in the more common case that we have more data than processors ($n > p$), we get good utilisation of the processors: a speedup of close to p , which is all that we can really expect, given our available hardware resources.

Rank-polymorphic array languages have historically been popular with their users because the human programmers like the expressiveness and clarity of the notation, without considering performance. But it ought to be true that such languages are well-suited to high-performance implementations on parallel hardware. (And this is our current research agenda.)

The bottom line is that we should think of Remora as a “map/reduce” architecture [1], but one where these parallel operators are notationally very lightweight: every function call comes with a notationally free map, and reductions are likewise easy to invoke. Thus, we can program these kinds of computations very easily, in a context of fine-grained parallelism,

Some basic uses of reduce

Here is the definition of the `vmag` function we’ve been using in our examples:

```
(define (vmag [v 1])
  (square-root (reduce/zero + 0 (square v))))
```

The function consumes vectors, hence the “1” rank of its `v` parameter. We first use the scalar `square` operator to produce a vector whose elements are the squares of the input vector’s elements. Then we sum these elements with a reduction

operator, and take the square-root of the result. Note that we did this without ever indexing into a vector or writing a loop.

To write the factorial function, we use the primitive `iota` function, which takes a vector specifying an array shape, and produces an array of that shape, whose elements are the naturals $0, 1, 2, \dots$ laid out in row-major order:

```
(iota [5])
[0 1 2 3 4]
(iota [2 3])
[[0 1 2]
 [3 4 5]]
(+ 1 (iota [5]))
[1 2 3 4 5]
(reduce/zero * 1 (+ 1 (iota [5]))) ; 5! = 120
120
(define (fact [n 0])
  (reduce/zero * 1 (+ 1 (iota [n]))))
(fact [0 3 5 10])
[1 6 120 3628800]
```

Some simple statistics

We can average the elements of a vector with this function:

```
(define (mean [xs 1])
  (/ (reduce + xs)
     (length xs)))
```

The `length` function is another function (like `append` and `reduce`) that consumes its entire argument as its cell; it returns the size of its argument's initial or leading dimension. Thus, applying `length` to a 3×5 array produces 3.

We can now define variance and covariance using `mean`:

```
(define (variance [xs 1])
  (mean (square (- xs (mean xs)))))

(define (covariance [xs 1] [ys 1])
  (mean (* (- xs (mean xs))
           (- ys (mean ys)))))
```

In variance, the subtraction operation uses principal-frame replication to subtract a scalar (the mean of the vector) from each element of the vector. The scalar square function is lifted to apply it pointwise to all the elements of its vector argument. Similarly, in covariance, the scalar multiply operation `*` is lifted to pointwise multiply the two argument vectors, producing a vector result, which is then averaged with `mean`. All of this is accomplished without needing to write an explicit loop or array index; instead of operating on scalar data, the program's operations are applied to entire collections.

One-dimensional convolution

We can convolve a vector of sample data `v` with a weighted window `w` in three lines of code:

```
(define (vector-convolve [v 1] [w 1])
  (reduce + (* (rotate v (indices-of w))
              w)))
```

The key to this function is the lifted `rotate` operation. The `rotate` function takes an array of rank `r` and a vector of `r` rotation amounts, one for each axis of the array. Thus, if we want to rotate a matrix `m` by 3 rows and 2 columns, we write `(rotate m [3 2])`. Here are some examples:

```
(rotate [2 3 4 5 11] [2])
[4 5 11 2 3]

(rotate [[2 3 4 5 11]
        [1 4 9 16 25]]
        [1 0]) ; Swap rows; don't rotate cols.
[[1 4 9 16 25]
 [2 3 4 5 11]]
```

When we rotate some data array `A` using an *array* of rotation vectors, Remora's rank-polymorphic lifting rules cause it to rotate `A` by each of the rotation vectors:

```
(rotate [2 3 5 7] ; Rotate this vector
        [[0] [1] [2]]) ; with 3 different rotations.
[[2 3 5 7]
 [3 5 7 2]
 [5 7 2 3]]
```

The principal frame of the operation is given by the rotation argument—since `rotate` consumes its entire first argument as its cell, the frame for the first argument is always a scalar frame, whose shape `[]` is always a prefix of the second argument’s frame. Here, each individual rotation produces a vector result; these three vectors are collected into the principal frame to produce the final matrix result.

The `indices-of` function takes an array of any rank r , and produces an array of rank $r + 1$. If we consider the resulting $r + 1$ -rank array to be an r -rank array of vectors, each of these vectors is its location in the containing array. Thus, if `m` is a 2×3 array, we get a $2 \times 3 \times 2$ result:

```
(indices-of m) ; m is a 2x3 matrix.
[[[0 0] [0 1] [0 2]]
 [[1 0] [1 1] [1 2]]]
```

The `indices-of` function is frequently useful, as it essentially reifies the iteration space of an array.

When we apply `indices-of` to the weights vector `w` of length n , we get an $n \times 1$ matrix `[[0] [1] ... [n - 1]]`. When we use this matrix for the rotation argument in the convolution code, we rotate the vector of sample data `v` by the rotation distances $0, 1, \dots, n - 1$. This produces the matrix

$$\begin{bmatrix} v_0 & v_1 & v_2 & \dots & v_{m-2} & v_{m-1} \\ v_1 & v_2 & v_3 & \dots & v_{m-1} & v_0 \\ v_2 & v_3 & v_4 & \dots & v_0 & v_1 \\ \vdots & & & & & \\ v_{n-1} & v_n & v_{n+1} & \dots & v_{n-3} & v_{n-2} \end{bmatrix}.$$

That is, the top row is the original sample vector; the second row is the sample vector rotated once; the third row is the sample vector rotated twice; and so forth. The width of the matrix is the length of the data vector `v`; its height is the length of the weight vector `w`; and each column of the matrix is one sample window’s worth of data. When we multiply this matrix by the weight vector, the rank-polymorphic lifting rules of Remora multiply the top row of the matrix by the first weight; the second row by the second weight; and so forth. After this, we simply sum each column of the result, collapsing the matrix vertically and producing the final convolution vector.

Again, note that we did not have to write explicit loops, nor did we ever need to use indexing to extract scalar values out of an array, instead operating on entire aggregates in parallel.

We leave it as an exercise for the interested reader to write a version of this function that performs a two- or three-dimensional convolution.⁵

Reranking gives control of frame/cell factoring

Remora's fixed frame-replication strategy sometimes doesn't do what we want. For example, if we have an n -element vector v and an $n \times n$ matrix m , we can add the first element of v to the first row of m , the second element of v to the second row of m , and so forth, very simply. The structure of the addition exactly matches the fixed architecture of Remora's principal-frame replication machinery, so we only need to write:

```
(+ v m)
```

However, suppose we want to add the first element of v to the first *column* of m , and so forth? We manage this by means of η -expanding the $+$ operation, adjusting the frame/cell split with the cell-rank parameter annotations on the wrapper λ term. Consider this example:

```
(define v [10 100])
(define m [[1 2]
          [3 4]])

((λ ([x 1] [y 1]) (+ x y))
 v
 m)
```

The key point here is that our η -expanded version of $+$ is *not identical* to our original $+$ function. We've shifted the function parameters from 0-rank cells to 1-rank cells, as marked by the 1 rank-annotations on the x and y formal parameters. This means that the new function cuts its arguments up into cells differently from the way that $+$ would do it. The principle frame is given by the two rows of the matrix argument, so the application is distributed over these rows and the single cell of the v argument is replicated across this distribution, giving the following sequence of execution steps:

⁵Have fun. The 2D case requires one additional function call; the 3D case, two more.

```

((λ ([x 1] [y 1]) (+ x y)) [10 100] [[1 2]
                                     [3 4]])
⇒
(((λ ([x 1] [y 1]) (+ x y)) [10 100] [1 2])
 ((λ ([x 1] [y 1]) (+ x y)) [10 100] [3 4]))
⇒
[(+ [10 100] [1 2])
 (+ [10 100] [3 4])]
⇒
[[(+ 10 1) (+ 100 2)]
 [(+ 10 3) (+ 100 4)]]
⇒
[[11 102]
 [13 104]]

```

... which is exactly what we wanted.

Manipulating the way Remora's frame-based distribution works with a re-ranking η -expansion is a standard idiom when programming in rank-polymorphic languages. One way to think of this is to bear in mind that function application, in a rank-polymorphic language, is a more complex mechanism than in the classic λ calculus. In some sense, every function application comes wrapped in its own set of nested loops. As we saw in this example, when the computation pushes an array argument from the site of the function application off to the body of the function being called, the argument is "cut up" into a collection of cells, and the function application is replicated in parallel across these cells. All of this implicit loop structure is why we don't have to write our own, explicit loops.

Programming in a rank-polymorphic language such as APL, J and Remora involves developing a reflexive understanding of how principal-frame cell replication causes arguments to be broken up and distributed. Because programmers frequently tune this mechanism with reranking, Remora provides a syntactic shorthand for doing so. We can write the η -expanded addition term from the above example with the reranking \sim notation: $(\sim(1\ 1)+\ v\ m)$. In general, writing

$$\sim(r_1 \dots r_n) \text{exp}$$

desugars to

$$(\lambda ([v_1\ r_1] \dots [v_n\ r_n]) (\text{exp}\ v_1 \dots v_n))$$

for fresh parameters v_i .⁶ That is, it permits us to specify the cell ranks r_i for the function's arguments.

Reranking is often useful in the context of the special functions that consume their entire actual argument as their cell, such as `append`, `rotate` and the `reduce/scan` family of functions (which effectively constitute a distinct component of Remora's control story).

For example, recall that `append` assembles its arguments together along their leading or initial dimension, so appending two matrices stacks one above the other. A re-ranked `append`, however, can assemble two matrices side-by-side:

```
(define m1 [[0 1]
            [2 3]])
(define m2 [[10 20]
            [30 40]])

(append m1 m2) ; m1 above m2
[[0 1]
 [2 3]
 [10 20]
 [30 40]]

(~(1 1)append m1 m2) ; m1 to the left of m2
[[0 1 10 20]
 [2 3 30 40]]
```

The reranked `append` works by distributing the `append` across the vector cells of the two arguments, assembling the results into a vector frame. After the reranked application, we have the intermediate frame-of-cells result:

```
[(append [0 1] [10 20])
 (append [2 3] [30 40])]
```

⁶This is almost true: it ignores the possibility that the evaluation of the function expression *exp* might have a side effect of some kind. In the presence of side effects and a call-by-value semantics, we must use the safer desugaring

```
(let ((f exp)) (λ ([v1 r1] ... [vn rn]) (f v1 ... vn)))
```

In practice, the correct, side-effect-safe desugaring almost always reduces to the more informal one we initially gave, as the function term being reranked is typically either a variable or a λ term.

Reranking is especially useful in the context of the `reduce/scan` family of operators. The default behavior of `reduce` is to collapse the array argument along its initial dimension:

```
(reduce + [[0 1 2] ; Add the first row to
          [0 10 100]]) ; the second row.
[0 11 102]
```

However, we can sum *across* the matrix by reranking the reduction to apply the reduction operation independently to each row of the input:

```
(~(0 1)reduce + [[0 1 2]
                [0 10 100]])
```

Here, the `reduce` application is distributed across the vector (rank 1) cells of the data matrix by the “1” element of the reranking, giving

```
[(reduce + [0 1 2])
 (reduce + [0 10 100])]
```

which produces final result

```
[3 110]
```

In this manner, we can use reranking to control the axes across which we iterate when performing a reduction on data collections.

Matrix multiplication

All of these computational mechanisms—rank polymorphism, principal-frame replication, reduction, and reranking—come together when we write the standard matrix-multiplication function from linear algebra. We begin by defining a function `v*m` that multiplies an n -element vector v times an $n \times p$ shaped matrix m , producing a p -element vector result. We want the first element of the result to be the dot product of v with the first column of m ; the second element of the result to be the dot product of v with the second column of m , and so forth:

```
(define (v*m [v 1] [m 2]) (reduce/zero + 0 (* v m)))
```

We're done: to multiply matrix a by matrix b, we simply apply v*m to the two matrices: (v*m a b)! The individual rows of a will be taken as the vector cells of the first argument, and each one independently multiplied by the entire matrix b, which will be taken as a single cell and replicated across the individual multiplies.

We can package this up with a definition that specifies rank-2 (that is, matrix) inputs as follows:

```
(define (m*m [a 2] [b 2]) (v*m a b))
```

...but note that this is just a reranked v*m, so could alternatively define the function this way:

```
(define m*m ~(2 2)v*m)
```

If two lines of code seems overly prolix, we can write matrix multiply in a single line of code by pulling all the cell/frame rank manipulation into the reranking notation:⁷

```
(define (m*m [a 2] [b 2])
  (~(0 0 2)reduce/zero + 0 (~(1 2)* a b)))
```

Polynomial evaluation three ways

At the beginning of this tutorial, we imagined a polynomial-evaluation function that takes a vector of coefficients, and an x value at which to evaluate the polynomial. We can define this function in several ways; the various definitions illuminate the considerations that apply to writing efficient, scalable code in Remora.

We begin with a straightforward definition:

```
;;; Simple polynomial evaluation
(define (poly-eval [coeffs 1] [x 0])
  (reduce/zero + 0
    (* coeffs
      (expt x (iota [(length coeffs)]))))))
```

⁷We like to think of Remora as a bit less terse and a bit more readable than classic APL...but we'll admit that this definition is pushing the boundaries of clarity for non-expert Remora programmers.

The innermost (`iota [(length coeffs)]`) term produces a vector of exponents. Suppose, for example, that the coefficients vector is length 4, with elements $[c_0 \ c_1 \ c_2 \ c_3]$. Then this inner expression produces (`iota [4]`), which is the vector $[0 \ 1 \ 2 \ 3]$. The exponentiation function `expt` raises `x` to all four of these powers, producing the vector $[x^0 \ x^1 \ x^2 \ x^3]$. We multiply this vector, point-wise, by the coefficients vector, and sum the result with a reduction, producing the final answer.

This definition is simple and clear, but we do a lot of redundant multiplication when we compute each power of x independently of the others. All told, our four-term polynomial example will do $0 + 0 + 1 + 2 = 3$ multiplies to compute the four powers of x that we need—that is, this code does a quadratic number of multiplies.

For a four-term polynomial, this is not much of a problem, but if our polynomial has a hundred terms, it is a significant waste of computation. We would rather do the exponentiations for our hundred-term polynomial using a total of 98 multiplies, instead of the 4,851 multiplies that this definition performs.

If we compute our polynomial with Horner's rule, that is, $c_0 + x(c_1 + x(c_2 + xc_3))$, we'll only require a linear number of multiplications. This gives us the following definition, which directly instantiates Horner's rule using a right-to-left fold along the coefficients vector:

```
;;; Efficient on serial processor
(define (poly-eval [coeffs 1] [x 0])
  (fold-right (\ ([coeff 0] [acc 0]) (+ coeff (* x acc)))
             0
             coeffs))
```

Unfortunately, using a fold operation makes it much harder to parallelise the code. Again, this is not particularly important if our polynomials are of low degree, but if we were evaluating large polynomials, we might want a function that can efficiently make use of multiple processors to execute in parallel. We can achieve this with our final definition:

```
;;; Efficient on serial or parallel processor
(define (poly-eval [coeffs 1] [x 0])
  (reduce/zero + 0
              (* coeffs
                (open-scan/zero * 1
                               (with-shape coeffs x))))))
```

This definition uses the `with-shape` function, which uses the elements from its second argument (the scalar `x`) to produce an array whose shape matches its first argument (the `coeffs` vector). If the data array has too few elements, `with-shape` cycles through them repeatedly; if the data array has too many elements, then trailing, unneeded items are silently discarded.

If our `coeffs` vector has n coefficients, the `with-shape` expression produces an n -element vector, whose elements are all `x`.⁸ We then use the `open-scan/zero` function to multiply together the elements of this vector. This scan operator is an external scan that produces a result with the same number of elements as the input; it does this by not producing the final element of the full scan. The result is a vector of exponentiations $[x^0 \ x^1 \ \dots \ x^{n-1}]$. Computing this with a scan has two advantages: we only do a linear number of multiplies, and we do so in a fully parallelised manner. From here, the code is straightforward: we multiply the vector of x -powers by the coefficients vector and sum the terms. Note that the summation is done with a reduction, so this part of the computation is also parallelisable. And, of course, the data-parallel bits of the computation expressed with basic rank-polymorphic frame/cell distribution are trivially parallelisable, as well. (In the case of this code, this is the point-wise multiplication of the coefficients by the x -powers vector.)

Conditional code in Remora

Conditional code in rank polymorphic languages such as Remora is done a bit differently, in order to express the computation in a parallel way. Remora provides the exact same `if` and `cond` forms that LISP and Scheme provide, but these forms are considered barriers to parallelism. So we can write a serial version of factorial as follows:

```
(define (fact [n 0])
  (if (zero? n) 1
      (* n (fact (- n 1)))))
```

⁸An alternate way to make a vector of `x` values with the same shape as the coefficients vector would be to map a constant `x`-producing function across `coeffs`, with the expression `((λ ([y 0]) x) coeffs)`, or, equivalently, with the `let` form of the β -redex:

```
(let ((y 0 coeffs)) x)
```

The previous version of factorial that uses `iota` and `reduce/zero` (page 18) permits the multiplications to be done in parallel.

More in line with the general data-parallel focus of the language, we can filter a collection of items by some criterion using the `filter` function. This code shows how to use `filter` to select out all the positive elements of a vector:

```
(define nums [0 5 -7 -22 91 100])

(filter (> nums 0) nums)
[5 91 100]
```

The `filter` function takes a boolean “selection” vector and a data array; the boolean vector must be as long as the initial dimension of the data array (so the data array must have rank greater than zero—it may not be a dimensionless scalar). If the boolean vector has d_1 items, and the data array has shape $[d_1, d_2, \dots, d_r]$, the function views the data array as a collection of d_1 sub-arrays, each with shape $[d_2, \dots, d_r]$; it selects out all the sub-arrays whose corresponding boolean is true.

So we can select certain rows of a matrix with

```
(filter [#t #f #f #t #t]
      [[ 0  1  2] ; yes
       [16 17 18] ; no
       [ 9 10 11] ; no
       [22 23 24] ; yes
       [96 97 98]]) ; yes
[[ 0  1  2]
 [22 23 24]
 [96 97 98]]
```

A similar function, `partition`, takes the same arguments and splits the data set into two collections, then uses Remora’s multiple-value return feature to return two distinct arrays.

To select columns from a matrix, rather than rows, we rerank the function so that it is individually applied to each row:

```
(~(1 1)filter [ #t #f #t]
             [[ 0  1  2]
              [16 17 18]
              [ 9 10 11]])
```

```
[[ 0 2]
 [16 18]
 [ 9 11]]
```

The simple function `select` takes a boolean and two other values, returning the first if the boolean is true, and the second, if it is false. This function is fundamentally defined on scalars, but it is frequently employed in a lifted capacity:

```
(select #t 3 4)
3
```

```
(select #f [8 1 2] [20 3 9])
[20 3 9]
```

```
(select [#t #f #f #t #t]
        [ 0 1 2 3 4]
        [20 21 22 23 24])
[0 21 22 3 4]
```

```
;; Merge a default value into a data collection
(select [#t #f #f #t #t]
        [ 0 1 2 3 4]
        100)
[0 100 100 3 4]
```

A generalisation of `filter` is the `replicate` function. Instead of taking a vector of booleans, it takes a vector of counts that specify how many copies of each item to provide:

```
(replicate [1 3 0 2]
           [20 73 99 14])
[20 73 73 73 14 14]
```

An example of using these data-parallel operators to write conditional code is a function that takes a string (that is, a vector of characters) and converts each contiguous run of spaces into a single space:

```

;;; Keep a char if it's not a space, or
;;; its left neighbor isn't a space.
(define (collapse-spaces [s 1])
  (let* ((ns (not (char=? #\space s)))
        (lns (drop-right1 (append [#t] ns) 1))))
    (filter (or ns lns) s)))

```

The `drop-right1` function takes a single drop count (hence the “1” in its name) which says how to shorten the leading dimension of its initial argument. (There is a more general `drop-right` that permits the programmer to provide a vector of drop counts, one for each dimension of the array.) In our example above, we are shifting the `ns` vector right by one element, filling in on the left with an initial true value (using `append`), and discarding the rightmost element (using `drop-right1`).

As usual, this function is completely parallelisable (since `filter` can be implemented using similar techniques to parallel reductions and scans); it is defined without recourse to iteration or array indexing.

Indexing items from arrays

Remora provides the ability to index into arrays. However, programs that use indexing to fetch a single element from an array at a time will not be able to operate on the array in parallel, so this is generally considered bad programming style, counter to Remora’s general data-parallel paradigm.

The proper way to use indexing is to employ it in a lifted manner: instead of using a *single* index to fetch a *single* value from an array, use an *array* of indices to fetch a *collection* of values from an array.

Let’s define a three-dimensional $3 \times 2 \times 4$ array to use for an example.

```

(define a [[[ 1  10 100 1000] ; plane 0
            [ 2  20 200 2000]]

           [[ 0  2  4  6] ; plane 1
            [ 1  3  5  7]]

           [[30 31 32 33] ; plane 2
            [40 41 42 43]]])

```

We can select the element in plane 1, row 1, column 2 with

```
(index a [1 1 2])
5
```

Note that the returned value is an array, of course—all expressions in Remora produce array results (in this case, a scalar array).

The `index` operation is just a function, so it lifts like any function in Remora. If we use a vector of indices (that is, a vector of index vectors, which is to say, a matrix), we will get a vector of results:

```
(index a [[1 1 2]
          [1 1 2]
          [0 1 3]])
[5 5 2000]
```

If we use a matrix of indices (that is, a rank-3 array), we will get a matrix of results:

```
(index a [[[1 1 2] [0 1 3]]
          [[2 0 0] [1 0 3]]])
[[5 2000]
 [30 6]]
```

The lifting rule for indexing is that the `index` function takes its entire first argument as its source-array cell, and breaks its second argument up into a frame of index vectors. However, we can rerank `index` to break the first, source-array argument up into smaller cells. So, to fetch the element at row 0 column 2 from every plane of `a`, we rerank `index` to fetch from matrices:

```
(~[2 1]index a [0 2])
[100 4 32]
```

The `index` function can select out an entire subarray from an array, if we provide it with an index vector that is shorter than the full rank of the array. To select plane 2 of `a`, we write

```
(index a [2])
[[30 31 32 33]
 [40 41 42 43]]
```

Likewise, we can select plane 2 row 1 with

```
(index a [2 1])  
[40 41 42 43]
```

Note that the `index` function considers an index to be a vector of integers. As a special case, the function `index-item` takes a scalar integer as index and uses it to select along the first dimension of the source array. So we can select plane 1 of array `a` with

```
(index-item a 1)  
[[0 2 4 6]  
 [1 3 5 7]]
```

Indexing and performance

Pumping a large collection of indices through an array is a powerful computational act: we can rotate, mirror, replicate, select, sort, and arbitrarily permute the elements of an array by performing various calculations on arrays of indices and then performing a single group index operation. This power comes at a cost, however: it is not likely that a compiler will be able to guarantee locality for the sequence of element fetches required, nor will it be able to guarantee statically that each index is a legal one, so the individual fetches will all have to be checked for safety.

So, it is best to regard Remora's lifted `index` operation as, in fact, a big *communications* step, that shuffles data from one layout scattered across the processors executing the program to a new layout; such a step taxes the computer's communications hardware, not its computational resources.

The more structured array manipulations performed by Remora's library of `rotate`, `mirror`, `take`, `drop`, *etc.* functions can provide better performance for both of these reasons (locality and elimination of index-safety checks).

Selecting regions of an array

An operation related to indexing is the family of subarray functions. While `index` produces a single scalar, these functions return an array whose rank is the same as the input array.

For example, we can select a $2 \times 2 \times 2$ cube out of array `a` starting at index `[1 0 2]` with

```

(subarray a [1 0 2] [2 2 2])
[[[4 6]
 [5 7]]

 [[32 33]
 [42 43]]]

```

The second argument to `subarray` is the *start* index; the third argument is the *shape* vector. The shape vector can be shorter than the rank of the source array; unspecified elements select the entire corresponding dimension of the array. Thus we can select out two planes of `a` starting at index `[1 0 2]` with

```

(subarray a [1 0 2]) [2])
[[[ 4 6]
 [ 5 7]]

 [[32 33]
 [42 43]]]

```

This gets us 2 planes beginning with plane 1; all rows beginning with row 0; and all columns beginning with column 2. Note that the initial elements of the shape of the result array are given by the shape argument (in this case, `[2]`).

The variant function `subarray/wrap` permits the programmer to select subarrays with toroidal topology, or wrapping on dimensions, if the selected subarray selects beyond the end of some dimension of the source array.

The variant function `subarray/fill` permits the programmer to specify a default “fill” value if the selected subarray extends beyond the end of some dimension of the source array.

Sorting

The `grade` function sorts data; it returns a permutation vector of integer indices that can then be used with `index-item` to shuffle the data.

```

(define v [3 1 4 1])
(grade < v)
[1 3 0 2]

```

```

      (index-item v (grade < v))
[1 1 3 4]
      (index-item v (grade > v))
[4 3 1 1]

```

The grade function provides a stable sort. The comparison function is applied to the “items” of the data array, that is, if the data array has rank r , it is applied to subarrays of rank $r - 1$.

So we can sort the rows of a matrix into ascending order, using a lexicographic sort to compare two rows, with

```

;;; Lexicographic order for two vectors
;;; The reduction picks out the leftmost non-zero
;;; element of its input vector v-w.
(define (lex< [v 1] [w 1])
  (negative? (reduce (λ (x y) (select (zero? x) y x))
                    (- v w))))

(define m [[3 1 4 1 5]
          [2 7 1 8 3]
          [1 6 1 8 0]])

(index-item m (grade lex< m))
[[1 6 1 8 0]
 [2 7 1 8 3]
 [3 1 4 1 5]]

```

If we just want the sorted data and don’t need the permutation index vector, we can directly use the sort function:

```

(sort < [3 1 4 1])
[1 1 3 4]

```

Explicitly typed Remora

The code examples we've seen up to this point have been written using a variant of the language that is *dynamically typed*, in the same way that Scheme, LISP, or JavaScript are. However, serious programming in Remora is done in a language that comes with a type system—one that is expressive enough to include descriptions of the ranks and shapes of the arrays on which functions operate. This means that a well-typed program comes with a guarantee that a large class of errors involving functions being applied to arrays with incompatible shapes cannot happen at run time.

The down side of the type system's complexity and expressiveness is that Remora's type system is quite verbose—in some cases, types can be larger than the terms they annotate. Remora is intended to be programmed by humans in a variant of the language that supports *type inference*: the programmer writes code with no (or few) type annotations, and a powerful inference algorithm fills in the missing type machinery, producing a fully typed term. The design and implementation of type-inference technology for Remora is the subject of current research [4] and beyond the scope of this tutorial.

Even in a type-inferred context, however, there are occasions where programmers may wish to include some explicit types in their programs (*e.g.*, at module boundaries), so the implicitly typed variant of the language includes the ability to mix explicitly typed and type-inferred code. Also, when type-inferenced code has array-shape bugs, these will be reported to the programmer as type errors (analogously to the way that type-inferred SML or OCaml code is handled). So even when programmers don't write types in their source code, types will still show up in the debugging process.

In this section, we'll introduce Remora's type system by means of a series of examples of steadily increasing complexity. Keep in mind, however, that while types in Remora are a powerful skeleton that provides Remora's iteration and parallelism control structure, these types are typically something that—like a human skeleton—are intended to operate hidden from view.

Types and indices

Remora's type system is a kind of (restricted) dependent type system, parameterised in two ways: over other types, and over “indices,” which represent array dimensions and shapes.

Splicing-shape notation	Core notation
[d1 @s 5 (+ 1 d2)]	⇒ (++) (shape d1) @ _s (shape 5 (+ 1 d2)))
[3 4]	⇒ (shape 3 4)
[@s1 @s2]	⇒ (++) @s1 @s2)
[@s]	⇒ @s

Figure 1: The splicing-shape notation is syntactic sugar for writing array shapes.

- A *dimension index* describes a single dimension of an array. A dimension can be a constant, literal dimension, such as 3 or 17; an index variable, such as d1 or size; or the sum of dimensions, such as (+ d1 7).
- A *shape index* represents a contiguous sequence of dimensions from an array. A shape can be given by listing a sequence of dimension indices in a shape form, e.g., (shape 50 d2 12). We can also write an index variable, such as @s1 or @dims, for an entire shape, or append shapes together with the ++ shape operator, e.g., (++) @s1 (shape 3 1)).

Note that the two kinds of index variable, dimension and shape, are orthographically distinguished: shape-index variables begin with an “@” character, while dimension-index variables do not.

Now that we’ve spelled out the core syntax for indices, we’ll shift to a syntactic sugar that is more readable: “splicing shape notation.” In a syntactic context where a shape can occur, a square-bracket delimited sequence of shape and dimension indices stands for the sequence of those indices, with shape indices “spliced” into the sequence. Figure 1 gives examples showing the correspondence between the splicing notation and the core notation.

Remora has two kinds of types. First, we have types that describe the individual elements of arrays, which includes basic primitives, such as `int`, `char` and `bool`, and type variables, such as `t1` or `e1t`. We can write the type of a function that takes arguments of (array) type τ_1, \dots, τ_n and produces a result of (array) type τ_{result} with the type

$$(\rightarrow (\tau_1 \dots \tau_n) \tau_{\text{result}})$$

The basic array type is written $(A \ t \ \text{shape})$, where t is the (element) type of the array’s constituent atoms, and shape is the shape of the array. So a 3×4 array of integers has type $(A \ \text{int} \ (\text{shape} \ 3 \ 4))$, while a boolean scalar has

type `(A bool (shape))`. Expressions can be polymorphic over array types; like shape-index variables, array-type variables must be written with an initial “@” character, *e.g.*, `@a1` or `@x`.

Thus, the type of the addition function is

$$(\rightarrow ((A \text{ int } (\text{shape})) (A \text{ int } (\text{shape}))) (A \text{ int } (\text{shape})))$$

That is, the addition function takes two arguments, each a scalar array (with empty shape `(shape)`) of `int` elements, and returns another scalar array of `int`.

Beyond this core syntax, we use the syntactic sugar

$$[t \ i_1 \ \dots \ i_n]$$

to mean the array type `(A t [i1 ... in])`. So we can write the type of the addition function a bit more succinctly as

$$(\rightarrow ([\text{int}] [\text{int}]) [\text{int}])$$

Finally, if we write an element type in a syntactic context where an array type is expected, it is taken to mean a scalar array of that type. So we can finally reduce the type of the addition function simply to

$$(\rightarrow (\text{int int}) \text{int})$$

However, this is all convenience syntax which should be taken to be shorthand for the original type of the three above.

We also have types for type-polymorphic expressions:

$$(\forall (tv_1 \ \dots \ tv_n) \ \tau)$$

Applying a value of this type to n types will produce a value of type τ , with the type variables tv_i instantiated to the given types. Note that the type variables tv_i can be both element-type variables, and array-type variables.

Likewise, we have a type constructor for describing types that are dependent on (static) array dimension and shape indices:

$$(\Pi (iv_1 \ \dots \ iv_n) \ \tau)$$

is a type τ abstracted over the index variables iv_i , which can be both dimension or shape indices.

Keyword	ASCII	Keyword	ASCII
λ	fn	\forall	forall
$\text{T}\lambda$	t-fn	Π	Pi
$\text{I}\lambda$	i-fn	Σ	Sigma
\rightarrow	->		

Table 2: Remora has seven core forms whose identifying keywords use non-ASCII Unicode characters; all have ASCII equivalents.

Arguments	Function form	Application form	Function type
Values	$(\lambda ([v \ \tau] \ \dots) \ e)$	$(f \ arg \ \dots)$	$(\rightarrow (\tau \ \dots) \ \tau)$
Types	$(\text{T}\lambda (tvar \ \dots) \ e)$	$(\text{t-app } f \ \tau \ \dots)$	$(\forall (tvar \ \dots) \ \tau)$
Indices	$(\text{I}\lambda (ivar \ \dots) \ e)$	$(\text{i-app } f \ i \ \dots)$	$(\Pi (ivar \ \dots) \ \tau)$

Table 3: Remora has three distinct kinds of functions or abstraction, depending on what kind of arguments are given to the function: values, types or indices.

Remora and ASCII

As we move into the details of the explicitly typed language, we’ll increasingly see non-Roman, “mathematical” characters, such as Π and \forall (as well as the λ keyword we saw in the simpler, dynamically typed language). There are seven such keywords in the language, as shown in Table 2; all seven have simple ASCII equivalents. We’ll consistently use the more attractive Unicode / super-ASCII keywords in this tutorial, but it is worth noting that Remora programs can easily be written in the least-common denominator character set of ASCII.

Explicitly typed functions

Remora provides three kinds of functions—abstractions over values, types and indices—and three corresponding forms for applying them. Table 3 shows the syntax of all six forms, and the type constructors for each kind of function.

An example of a simple, monomorphic value abstraction is a function that takes a scalar integer and doubles it:

```
(λ ([i int]) (+ i i))
```

which desugars (as an expression) to the scalar array

```
(array []
  (λ ([i (A int (shape))])
    (+ i i)))
```

The function has type $(\rightarrow (\text{int}) \text{int})$. We can give the function a name with a `define` form

```
(define double (λ ([i int]) (+ i i)))
```

which can also be written as a “functional” definition with the syntactic sugar

```
(define (double [i int])
  (+ i i))
```

We apply a function to arguments simply by putting the function-producing expression and the argument-producing expressions in a form with no keyword at all. Thus, the expression

```
(double [3 2])
```

applies our doubling function to a vector. Remora’s lifting semantics maps the scalar-consuming doubler across a vector frame, producing the result `[6 4]`.

Another example showing the application of a value-consuming function is the body expression of our doubler function, `(+ i i)`. The function-producing subexpression `+` evaluates to a scalar array whose single element is the addition function. The two argument-producing expressions `i` and `i` each produce a scalar integer array (as determined by the type specified for `i`). The function is then applied to the two integer scalar arguments.

We can write functions that are polymorphic—that is, abstracted over *type* arguments instead of value arguments—with the form $(\text{T}\lambda (tvar \dots) body)$. A very simple example is the polymorphic identity function

```
(define identity (Tλ (@t)
  (λ ([x @t]) x))
```

which has type

```
(∀ (@t) ( $\rightarrow$  (@t) @t))
```

We must apply this function to a type with a `t-app` form before we can apply it to an actual array:

```
((t-app identity [bool 2 3]) [[#f #t #f]
                               [#t #t #t]])
```

This code first applies `identity` to an array type which specialises `@t` to boolean matrices with two rows and three columns; the result is an identity function

```
(λ ([x [bool 2 3]]) x)
```

which can only be applied to 2×3 matrices of booleans. (Recall that the array type `[bool 2 3]` is syntactic sugar for the type `(A bool (shape 2 3))`.)

Note that type abstractions can have type-variable parameters that represent element types (written without a leading at-sign: `t`, `x`, *etc.*), or full array types (written with a leading at-sign: `@t`, `@x`, *etc.*).

As another example, consider the polymorphic function that is a “function doubler”—that is, it takes a function f and returns the function which applies f twice to its input. This function can be defined and given a name with

```
(define fdouble (Tλ (@t)
                   (λ ([f (→ (@t) @t)])
                     (λ ([x @t]) (f (f x))))))
```

It has type

```
(∀ (@t) (→ ((→ (@t) @t))
            (→ (@t) @t)))
```

As with our `identity` function, using `fdouble` requires a double application. First, we use a `t-app` to specialise the polymorphic function to a given type (such as floating-point scalars), then we apply the resulting function to the function we wish to double (in the following example, the sine function):

```
(define sine-twice ((t-app fdouble [float]) sin))

(sine-twice [[3.14159 2.71828]
            [0.0      1.0     ]])
```

Because `sine-twice` is typed to operate on scalar arrays, we are allowed to apply it to a matrix of floating-point elements by means of Remora’s rank-polymorphic function-call machinery—the 2×2 array is treated as a matrix frame of scalar cells, so we apply the sine function twice to each of the cells.

Finally, Remora permits the programmer to write functions that abstract over shape and dimension indices. Our `dot-product` function is an example of a dimension-polymorphic function: it operates on two arguments of fixed rank (they are both rank-1 vectors). The two arguments can be of any length, but must both be of the same length. Therefore, the type of `dot-product` is parameterised over the dimension index `len`, which represents the length of the two vectors

```
(Π (len)
  (→ ([int len] [int len])
      int))
```

This means that when we use `dot-product`, we must first apply it (using the `i-app` index-application form) to a dimension, and then apply the resulting value to the actual vectors:

```
((i-app dot-product 3) ; Specialise function for 3-vectors,
 [8 1 2]               ; then apply to
 [2 0 9])              ; two such vectors.
```

In this example, the `3` is not the *value* three—something we could multiply times four, or square to make nine. It occurs in the grammatical context of an *index*, so it denotes the *array dimension* three.

As another example, consider the `append` function, which appends two arrays of identical rank along their principal, or initial, dimension. For example, if we append a $2 \times 3 \times 5$ array of booleans and a $7 \times 3 \times 5$ array of booleans, we get a result which is $9 \times 3 \times 5$. So, the type of the `append` function is

```
(Π (da db @rest)
  (∀ (t)
    (→ ([t da @rest] [t db @rest])
        [t (+ da db) @rest])))
```

In our example above,

- index `da` would be 2, the initial dimension of the first argument,
- index `db` would be 7, the initial dimension of the second argument,
- index `@rest` would be `[3, 5]`, the remaining shape of both arguments, and
- type `t` would be `bool`, the type of the arrays' elements.

Here is an expression that uses `append` to concatenate two arrays `a1` and `a2` of the given shapes:

```
((t-app (i-app append 2 7 [3 5])
        bool)
   a1 a2)
```

This triple index/type/value application is a pattern that frequently occurs when using shape-polymorphic functions.

The types of scan and reduce functions

The `scan` / `reduce` / `fold` / `trace` family of functions are all both type- and index-polymorphic. In the terminology of these functions, we think of the source array as a vector of subarrays, called the “items” of the reduction, to be combined by the reduction operator. Thus, if we want to add together the planes of a 3-D array, the items of the array are its individual planes, each a 2-D matrix.

As a convenience, the types for these functions include an extra shape-index parameter that permits us to “pad” a “sub-item” operator like `+` when we want to use it to combine together higher-rank items. The full type for `reduce` is

```
(Π (d-1 @item-pad @cell-shape)
  (∀ (t)
    (→ ( (→ ( [t @cell-shape] [t @cell-shape] )
              [t @cell-shape])
          [t (+ d-1 1) @item-pad @cell-shape] )
        [t @item-pad @cell-shape])))
```

Suppose, for example, that we wish to sum the planes of a 3-D array `A` with shape `[5 7 4]`. We want to use the scalar `+` directly, rather than have to “up-rank” the operator to work on 7×4 matrices by tediously writing out the η -expanded

```
(λ ([x [int 7 4]] [y [int 7 4]])
  (+ x y))
```

This is the purpose of the `@item-pad` shape in the type of `reduce`: the operator we use to reduce the input array can operate on sub-items of the array, of shape `@cell-shape`; it will be padded up to the full items of the array with the extra dimensions of `@item-pad`.

So, in our $5 \times 7 \times 4$ array example, we have $d-1 = 4$, which means the initial or “iteration” dimension along which we reduce is $(+ \ d-1 \ 1)$, or 5, as is specified on line 4 of the type. Adding one in the type is how we require this dimension of the array to be at least one: reduce cannot operate on arrays where the iteration dimension is zero. (The intent of choosing “d-1” for the name of the index variable is that it be read as “d minus 1.”)

Continuing through the parameters of our type, the shape index `@cell-shape` provides the shape of the arguments consumed and produced by the reduction’s combining operator. Since our combining operator `+` works on scalars, we have `@cell-shape = []` and type `t = int`. However, we will actually be adding together 7×4 matrices, not scalars, so we must pad the shape up by specifying `@item-pad = [7 4]`.

Thus, to sum the planes of our $5 \times 7 \times 4$ array `A`, we write

```
((t-app (i-app reduce 4 [7 4] [])) int) + A)
```

The types of the other reduction, scan, fold and trace operators are similar in structure to the type of reduce, *mutatis mutandis*.

Boxes and existential types

“Boxes” are a mechanism in Remora that permit the language to handle two issues:

- **Array dimensions are not always statically known**

Remora’s index language gives programmers a limited ability to compute the shapes of arrays we are calculating from the shapes of the arrays that form the inputs to our calculations, using the `+` operator to add dimensions, and the `++` operator to append shapes.

Sometimes, however, the shape of some computed array is too complex to be described by such a restricted facility. Consider the function (written in the dynamically typed variant of the language):

```
(λ ([n 0]) (iota [n]))
```

Trying to come up with a typed version of this function runs into a problem: the length of the vector produced by the function is dependent on a *computed value*, not some other array shape.

Reasoning about the possible values that `n` might have is equivalent in decidability to the halting problem, so we have no hope of being able to solve this in general. Furthermore, Remora permits I/O, so the value of `n` could be supplied at run time by a human sitting at a keyboard, thus rendering it completely unknowable at compile time.

- **Some data collections are not regular arrays**

In Remora, we can write a vector of characters, *e.g.*, the five vowels

```
[\a \e \i \o \u]
```

more compactly as `"aeiou"`. So `["abc" "xyz"]` is a matrix with two rows and three columns.

Now, suppose we wish to represent the five days of the work week as character data. We cannot write this as a matrix of characters

```
["Monday"      ; *Not* a legal  
 "Tuesday"    ; expression!  
 "Wednesday"  
 "Thursday"  
 "Friday"]
```

This is not a legal array—the third row (Wednesday) has nine columns, while the first row (Monday) has only six. We could consider padding short rows with some kind of null or blank filler, but this is not a general solution to the problem of “ragged” data.

In Remora, a “box” is a thing that packages up an entire array as a single *element* in some other array. It is a mechanism that lets us regard a *collection* (an array) as a member (an element) of some other collection. Boxes package up two things together: a *value* (that is, an array), and a set of *indices* that we will need later when we *unbox* the array.

We can write our days-of-the-week example using boxes as

```
(boxes (len) [char len] [5]  
 ((6) "Monday"  )  
 ((7) "Tuesday" )  
 ((9) "Wednesday")  
 ((8) "Thursday" )  
 ((6) "Friday"  ))
```

Let's begin with the type of this expression, which is

$$[(\Sigma (\text{len}) [\text{char } \text{len}]) 5]$$

This is the type of vectors of length five whose elements each have item type

$$(\Sigma (\text{len}) [\text{char } \text{len}])$$

In English, this element type means, "There is some dimension len such that this box contains an array of type $[\text{char } \text{len}]$ —a character vector of length len ." What's more, each box in our vector of five boxes can use a *different dimension* for len . In one box, len can be five; in another, it can be twenty-seven; and so on.

The `boxes` form shows how we can build our example boxed array:

- The first part of the `boxes` form is a list of index variables over which the boxed-up arrays are existentially abstracted. In our case, we are only abstracting over the length of each boxed-up vector, so we only have the dimension index len .
- The second part of the form, $[\text{char } \text{len}]$, is the abstracted type: each boxed-up array must be a character vector of length len for some value of len .
- The third part of the form, $[5]$, is the shape of the array we are constructing: it is a five-element vector.
- Finally, we have a series of index/expression clauses. Each clause specifies a particular dimension for len and an expression that has type $[\text{char } \text{len}]$ *given that particular* len . Thus, the specific dimension is called the "witness" for the element type $(\Sigma (\text{len}) [\text{char } \text{len}])$.

In our example, "Monday" is a character vector of length six, so we specify $\text{len} = 6$ in the first clause; likewise, we specify $\text{len} = 8$ in the "Thursday" clause; and so forth.

The general syntax of such an array expression is

$$(\text{boxes } (\text{ivar } \dots) \tau [\text{dim } \dots] \\ ((\text{index } \dots) \text{exp}) \dots)$$

The $(ivar \dots)$ part of the form is a list of index variables over which each box's array type is existentially abstracted—in our example, we only needed one such index variable. The τ is an array type which gives the type of the array inside each box, which may refer to the $ivar$ index vars. The shape $[dim \dots]$ is the shape of the box array we are constructing. The rest of the form is a sequence of indices/expression pairs $((index \dots) \ exp)$. The form must have one of these pairs for each element of the array, just as with an array form. Each expression exp must produce an array of type τ once the witness indices are plugged into the type for the corresponding $ivar$ index variables.

As another example, here is a three-element vector of boxes, where each element is a boxed integer matrix:

```
(boxes (r c) [int r c] [3]
  ((2 2) [[1 2]
          [3 4]]))

((1 3) [[10 100 1000]])

((3 1) [[10] [100] [1000]]))
```

The first element of this vector is a 2×2 matrix; the second, a 1×3 matrix; and the third, a 3×1 matrix.

Boxes are frequently created one at a time, from some given array, so the syntactic sugar $(\text{box } ((ivar \ index) \dots) \ \tau \ exp)$ is a convenient shorthand for such a scalar box array; it is equivalent to

```
(boxes (ivar ...)  $\tau$  [] ; The "[]" means a scalar array;
  ((index ...) exp)) ; here is its one item.
```

For example, the expression

```
(box ((len 3)) [int len] [8 23 0])
```

boxes up a three-element integer vector, hiding the length of the vector, as the single item of a scalar array of boxes.

We can unpack an array of boxes with the `unbox` form. For example, suppose our days-of-the-week vector above is named `weekdays`, and we wish to compute a vector of booleans specifying which weekdays are exactly six letters long. We write

```
(unbox weekdays (day len)
  (= 6 ((t-app (i-app length len []) char) day)))
```

```
[#t #f #f #f #t]
```

For each box in the vector `weekdays`, the `unbox` form opens up the box; extracts the array and binds it to the variable `day`; extracts the corresponding dimension witness—which, in this example, is the length of the character vector `day`—and binds it to index variable `len`; and finally evaluates the body of the form. Each time the `=` numeric comparison is evaluated, it produces a boolean scalar array; these arrays are collected into the vector frame given by `weekdays`, and we get a five-element vector of booleans, informing us that only “Monday” and “Friday” have exactly six characters.

This explicitly typed code is somewhat ugly because the `length` function is type and shape polymorphic. We have to instantiate it with an `i-app` / `t-app` pair before applying to its actual value argument `day`. The full type of `length` is

$$(\Pi (d1 @s) (\forall (t) (\rightarrow ([t d1 @s]) int)))$$

(Note that `length` requires its argument to have rank greater than zero—that is the point of splitting out the initial dimension `d1` of its argument.) The type-inferred version is easier to read:

```
(unbox weekdays (day len)
  (= 6 (length char) day))
```

Note that we cannot pass `len` as a parameter to `=` because it is an index variable, not a run-time integer value. (Effectively, it is the business of `length` to shift the array’s length from the index world down to the value world.)

Boxes and arrays with dynamic shape

When we program in Remora, array shapes are statically known much of the time. For example, if we use `reduce` to collapse vertically an array with type `[char 3 5]` (that is, a 3×5 matrix of characters), then we know the result will have type `[char 3]`.

Sometimes, however, we need to compute an array whose shape is determined by some complex computation that cannot be expressed in terms of the shapes of the computation’s input values using the restricted notation of the Remora type

system (which only permits the programmer to append shapes and add dimensions). Other times, an array's shape may depend on values that are supplied only at run time, perhaps as input from the user running the program.

As described earlier, a standard example of this is `iota`. Because this function produces an array whose shape cannot be determined at compile time, it is defined to return this array as a scalar box, packaged up with its shape. Thus, the type of `iota` is

$$(\Pi (r) (\rightarrow ([\text{int } r]) (\Sigma (@s) [\text{int } @s])))$$

In English, this means that `iota` takes as its argument an integer vector of length `r` which specifies (as a run-time value, *not* as a compile-time shape index) the shape of the array to produce, and returns a scalar box containing an integer array of some shape `@s`. In fact, the shape `@s` of this array is the same as the input argument... except that the former is a shape index in our type system, and the latter is an actual value.

Note that when we open up a box made by `iota` with an `unbox` form binding `a` to the boxed array and `@s` to its shape

```
(unbox ((i-app iota 2) [r c])
      (a @s)
      ... a ... @s ...)
```

we know absolutely nothing, at the type-system level, about the shape of `a`. Even though a human can tell from the above code that `a` is a matrix with `r` rows and `c` columns, all that is known at the type/index level is that `a` *has* a shape (named `@s`), which is not very informative.

Remora provides a small family of restricted variants to the `iota` function, specialised to different ranks, which allow the type system to extract more information about the shape of the result: `iota0` through `iota9`. For example, `iota1` produces a (boxed) vector, while `iota2` produces a (boxed) matrix. Thus, the type of `iota1` is

$$(\rightarrow (\text{int}) (\Sigma (\text{len}) [\text{int } \text{len}]))$$

This function produces a scalar box that packages up an integer vector.

We can use `iota1` to write the factorial function:

```

(define (fact [n int])
  (unbox (iota1 n) (factors-1 len)
    ;; RED is type & index-specialised REDUCE/ZERO:
    (let ((red (t-app (i-app reduce/zero len [] []) int)))
      (red * 1 (+ 1 factors-1))))))

```

The purpose of the `let`-binding is simply to break up the code for clarity; it is not strictly necessary.

Avoiding `iota` and rank-monomorphic programming

The `iota` function is frequently used in (dynamically typed) APL and J to generate arrays that encode “iteration space,” as collections whose items specify locations in a sequence. When possible, programmers in Typed Remora are better off generating such arrays using the `indices-of` function, as the shape of the result can be determined by the type system, at compile time, and no boxing is necessary. See the one-dimensional convolution code on page 19 for an example of `indices-of` being used in this way.

Some functions in Remora that manipulate array shapes produce results whose shapes are difficult to express in terms of the shapes of their input arrays, given the bounded expressiveness of Remora’s type system. One example is `indices-of`. The final *dimension* of its result array is the *rank* of its input array: given a 7×10 matrix, it will produce a $7 \times 10 \times 2$ result; given a rank-3 array with dimensions $11 \times 5 \times 14$, it will produce an $11 \times 5 \times 14$ array of 3-vectors, that is, a rank-4 array with dimensions $11 \times 5 \times 14 \times 3$. Thus the type of `indices-of` is

```

(Π (@s)
  (∀ (t)
    (→ ([t @s])
      (Σ (r) ; r = (length @s)
        [int @s r])))))

```

We have to use a box type here because the type system is not expressive enough to say that the final dimension `r` of the result array is the length of the input array’s shape `@s` (that is, dimension `r` is the rank of the input array). The fact that `r` is the length of shape `@s` is, alas, not expressed in the types, but rather hidden away in a comment.

True rank-polymorphic programming is rare in Remora: programmers almost always know the ranks of their arrays. Because of this, we can work around the

limits of the type system by providing rank-monomorphic variants of functions such as `indices-of`. Thus, we have the nine functions `indices-of/1` through `indices-of/9`, each specialised to inputs of the specified rank. For example, `indices-of/2` is restricted to operate on matrices only; it has type

```
(Π (d1 d2)
  (∀ (t)
    (→ ([t d1 d2])      ; input d1 x d2 matrix
        [int d1 d2 2])))
```

Note that this function, happily, does not produce a boxed result; the result array's shape is completely known given the shape of the input.

If we use a rank-monomorphic version of `indices-of`, the fully typed version of the one-dimensional convolution given earlier is

```
(define vector-convolve
  (I λ (vlen wlen)
    (λ ([v [int vlen]] [w [int wlen]])
      ;; Specialise REDUCE & INDICES-OF/1
      ;; to our specific needs:
      (let ((red (t-app (i-app reduce wlen [vlen] [])
                        int))
            (io/1 (t-app (i-app indices-of/1 wlen)
                         int)))
        ;; The actual convolution:
        (red + (* (rotate v (io/1 w)) w))))))
```

Types in a Remora program constitute an “explanation” (or proof) to the compiler of the shapes of the values being computed by the program. Programming in Typed Remora requires programmers to develop a sense of how to expose and exploit this shape information.

Conditional code and boxes

In Typed Remora, functions such as `filter` and `replicate` (see page 27) necessarily produce boxed results, as there is no way at compile time to determine the size of the result. Thus, the type of `filter` is

```

(Π (da @s)
  (∀ (t)
    (→ ( [bool da] [t da @s] )
        (Σ (db) [t db @s])))

```

That is, `filter` takes a boolean vector of length `da`, and an array whose initial dimension is `da` and whose other dimensions are represented by the shape `@s`. We think of this array as a collection of `da` items, each of which is an array of shape `@s`.

The function uses the boolean vector to filter the collection; we retain each subarray whose corresponding boolean is the true value. These items are collected into an array of type `[t db @s]`, where `db` is the number of items that survived the filtering process. Since there is no way to determine the dimension `db` at compile time, it is hidden inside a scalar box.

If we repeat our previous example (page 28) filtering the rows of a 5×3 matrix, in a typed setting, we get the following code. The filter operation views the matrix as a collection of five subarrays (`da = 5`), where each subarray is a three-element vector (`@s = [3]`), whose elements are integers (`t = int`). Three of the five rows are selected by the filtering operation, so `db = 3`; as this cannot be determined at compile time, the resulting collection is packaged up in a box whose type abstracts away the dimension `db`. All the existential Σ type tells us about the result is that there is *some* dimension `db` giving the number of rows in the $db \times 3$ matrix.

```

((t-app (i-app filter 5 [3]) int)
  [#t #f #f #t #t]
  [[ 0  1  2]    ; yes
   [16 17 18]   ; no
   [ 9 10 11]   ; no
   [22 23 24]   ; yes
   [96 97 98]] ; yes

(box ((db 3)) [int db 3]
  [[ 0  1  2]
   [22 23 24]
   [96 97 98]])

```

References

- [1] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, January 2008.
- [2] Kenneth E. Iverson. *A Programming Language*. John Wiley & Sons, Inc., New York, NY, USA, 1962.
- [3] Jsoftware, Inc. Jsoftware: High-performance development platform. <http://www.jsoftware.com/>.
- [4] Justin Slepak. *A Typed Programming Language: The Semantics of Rank Polymorphism*. PhD thesis, Khoury College of Computing Sciences, Northeastern University, Boston, Massachusetts, 2020. In preparation.
- [5] Justin Slepak, Olin Shivers, and Panagiotis Manolios. An array-oriented language with static rank polymorphism. In *Proceedings of the 23rd European Symposium on Programming (ESOP'14)*, volume 8410 of *Lecture Notes in Computer Science*, pages 27–46, Grenoble, France, April 2014. Springer.
- [6] Justin Slepak, Olin Shivers, and Panagiotis Manolios. The semantics of rank polymorphism. *CoRR*, abs/1907.00509, 2019. Preprint of article currently in journal submission. <http://arxiv.org/abs/1907.00509>.