

Modular GPU Programming with Typed Perspectives

MANYA BANSAL, Massachusetts Institute of Technology, United States of America

DANIEL SAINATI, University of Pennsylvania, United States of America

JOSEPH W. CUTLER, University of Pennsylvania, United States of America

SAMAN AMARASINGHE, Massachusetts Institute of Technology, United States of America

JONATHAN RAGAN-KELLEY, Massachusetts Institute of Technology, United States of America

To achieve peak performance on modern GPUs, one must balance two frames of mind: issuing instructions to individual threads to control their behavior, while simultaneously tracking the convergence of many threads acting in concert to perform *collective operations* like Tensor Core instructions. The tension between these two mindsets makes modular programming error prone. Functions that encapsulate collective operations, despite being called per-thread, must be executed cooperatively by groups of threads.

In this work, we introduce Prism, a new GPU language that restores modularity while still giving programmers the low-level control over collective operations necessary for high performance. Our core idea is *typed perspectives*, which materialize, at the type level, the granularity at which the programmer is controlling the behavior of threads. We describe the design of Prism, implement a compiler for it, and lay its theoretical foundations in a core calculus called Bundl. We implement state-of-the-art GPU kernels in Prism and find that it offers programmers the safety guarantees needed to confidently write modular code without sacrificing performance.

1 Introduction

CUDA [39] is a low-level, imperative programming language for NVIDIA GPUs. These GPUs are organized into a hierarchy of compute resources. *Threads* are the basic unit of sequential execution, *blocks* are groups of threads that can cooperate through shared scratchpad memory, and the *grid* is the full collection of blocks launched for a computation. A GPU program executes in parallel across this hierarchy, but is written *from the perspective of a single thread*.

While operations are specified per-thread, some are only valid when executed *collectively* by a group of threads. The `__syncthreads()` intrinsic, for instance, synchronizes all threads within a block, and it will cause a deadlock if executed by only some of those threads. There are many such collective intrinsics, including Tensor Core instructions [41–43], warp shuffle operations [22], and other kinds of synchronization primitives [21]. These operations require programmers to carefully marshal compute resources to coordinate which threads execute which lines of code. As a result, such collective operations break the illusion of threads executing independently.

The conceptual clash between regular statements (executed by a single thread) and collective operations (executed by a group of threads) impacts not only hardware intrinsics, but also *user-defined* collective operations, like functions. While functions are invoked individually by threads, they may encapsulate collective behavior, creating a contradiction between the per-thread syntax of their invocation and the cooperative semantics of their execution.

This contradiction puts modularity at odds with correctness and is apparent even in widely used CUDA libraries that package common functionality via function interfaces [13, 38, 51]. Consider the following snippet of documentation taken directly from CUB [38], a library of parallel primitives. It describes the `BlockReduce` function [11], in which all threads in a block collaboratively apply a reduction operation, such as a maximum or prefix sum, over an array:

Computes a block-wide reduction for `thread0` using the specified binary reduction functor.

- The return value is undefined in threads other than `thread0`.
- A subsequent `__syncthreads()` threadblock barrier should be invoked after calling this method if the collective's temporary storage (e.g., `temp_storage`) is to be reused or repurposed.

The snippet above attempts to convey several assumptions implicit in `BlockReduce`'s implementation. First, the function is only well defined when invoked by all threads in a block. Second, because

it accesses memory shared among threads, any subsequent reuse of that storage requires synchronization to ensure all threads have completed their accesses. In effect, CUB attempts to retrofit CUDA with information about the compute and memory requirements of collective operations. By prefixing functions with identifiers such as **Block**, CUB creates ad-hoc “namespaces” for different functions that assume similar invariants. However, without a type system to statically enforce these invariants, correct usage of this function—and collective operations in general—depends on carefully reading and interpreting the documentation.

In this work, we ask: can we provide low-level control over collective operations while statically guaranteeing that they execute with the necessary compute resources? By reifying configurations of compute resources as type-level *perspectives*—so named because they describe the view of GPU resources against which each statement is written—we find that we can. Our key insight is that GPU programmers naturally map computations onto different compute resources, and, unlike CUDA, which obscures this mapping, we can track it with a type system. This tracking enforces that collective operations are executed with the necessary resources, while still allowing users to access low-level intrinsics. Further, it allows users to define and compose their own collective operations by specifying the perspective with which their code must be run.

Our approach departs from previous work, which attempts to resolve the mismatch between per-thread syntax and collective execution by restricting access to low-level operations. Tile-based languages like Triton [54], Helion [44], and Tilus [26], limit the user’s perspective to the block level, which prevents them from writing the highest-performance kernels. A variety of functional languages, [7, 27, 32] provide compile-time guarantees through their type systems but do not expose low-level control over hardware. Some other efforts, like Descend [35], aim to provide a low-level, memory-safe GPU systems programming language, but lack support for modern GPU features (like Tensor Cores or asynchrony) entirely. As a result, despite its flaws, CUDA remains the de-facto standard for writing high-performance kernels on modern GPUs [23].

We introduce Prism, a new low-level GPU programming language that enforces that operations are only executed with the correct view of hardware resources, enabling fearless composition. Inspired by dependency calculi [1], Prism statically tracks the configurations of compute and memory resources with type-level perspectives. We also develop Bundl, a core calculus underpinning Prism, provide formal rules for manipulating the perspectives of both code and data, and prove *type-and-perspective safety*. This ensures that Bundl is sound, and that code always has the right perspective to execute operations at run time. A parallel goal, alongside safe composition, is performance. We incorporate modern GPU features such as Tensor Cores and asynchronous data movement into Prism, and demonstrate that Prism can achieve the same performance as hand-written, highly optimized code on an H100 and a 4070 Ti Super. Our contributions are:

- (1) Prism, a low-level GPU language that tracks the logical grouping of compute and memory resources with type-level perspectives, empowering users to write modular code (Section 3);
- (2) Bundl, a formal model of Prism that tracks perspectives in its type system and operational semantics, along with a soundness theorem guaranteeing that programs execute operations only when they have been statically proven to possess the necessary perspective (Section 4); and
- (3) An implementation of Prism (Section 5) that demonstrates that it can support modern GPU features like Tensor Cores and asynchronous data movement, achieving performance comparable to hand-optimized CUDA implementations (Section 6).

Section 7 discusses related work, and Section 8 discusses the limitations of our approach and outlines future work.

```

1 int tid = threadIdx.x;
2 if (tid >= 0 && tid < 32){
3     float A[4], B[2];
4     float C[4] = { 0 };
5     // Populate A with unique values.
6     for (int i = 0; i < 4; i++)
7         A[i] = tid * 4 + 1;
8     // Populate B with unique values.
9     for (int i = 0; i < 2; i++)
10        B[i] = tid * 4 + 1;
11    // Issue a warp-level Tensor Core
12    // operation: D = A * B + C
13    // (eliding some typecasts).
14    asm("mma.sync.aligned.m16n8k8...""
15        "{%0, %1, %2, %3},    /*D*/"
16        "{%4, %5, %6, %7},    /*A*/"
17        "{%8, %9},           /*B*/"
18        "{%10, %11, %12, %13};/*C*/"
19        : "=r"(C[0]), "r"(C[1]), ...
20        : "r"(A[0]), "r"(A[1]), ...
21        : "r"(B[0]), "r"(B[1]),
22        : "r"(C[0]), "r"(C[1]), ...);}

```

Fig. 1. Warp-level Tensor Core instruction in CUDA.

```

1 tid : int @ thread[1] = id();
2 with group(thread[32]):
3     A : float[4] @ thread[1]
4     B : float[2] @ thread[1]
5     C : float[4] @ thread[1]
6
7     # Populate A with unique values
8     for i in range(0, 4, 1):
9         A[i] = tid * 4 + 1
10        C[i] = 0
11    # Populate B with unique values
12    for i in range(0, 2, 1):
13        B[i] = tid * 4 + 1
14
15    # Issue Tensor Core op.
16    intrinsic.mma(
17        A[0], A[1], A[2], A[3],
18        B[0], B[1],
19        C[0], C[1], C[2], C[3],
20        out=[A[0], A[1], A[2], A[3]])

```

Fig. 2. Warp-level Tensor Core instruction in Prism.

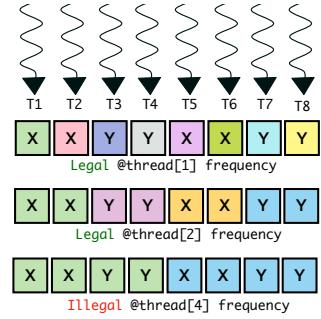


Fig. 3. The `@` syntax represents the frequency at which a variable varies across threads T1–T8. Each `@` denotes a coloring of the variable; threads that “view” the variable with the same color must observe the same value.

2 Background & Motivation

Before diving into the design of Prism, we begin with an overview of the GPU’s compute and memory hierarchies and outline the challenges posed by reasoning about them collectively. We also discuss how Prism can help solve these problems.

2.1 Compute Hierarchy

In CUDA, programmers launch computations that run on thousands of threads. These threads are organized into a *compute hierarchy* that defines how work is distributed and scheduled on the GPU. At the top of this hierarchy is the *grid*, representing all threads launched as part of a single computation. The grid is divided into *blocks*, each containing a user-specified number of threads. *Threads* are the finest unit of execution and the machine’s basic unit of sequential control.

Users describe a computation by writing a single program which is replicated across all threads. This program controls the behavior of individual threads by reading built-in identifiers like `threadIdx.x`—to determine a thread’s position within a block—and `blockIdx.x`—to determine the block’s position within the grid—at runtime. By making control flow decisions based on these two variables, users can assign each thread to its share of the full computation.

A natural and tempting way to interpret these built-in variables is to think of them as the indices of implicit “parallel-for loops” surrounding the program, where each iteration executes simultaneously. While this view is sufficient to understand CUDA’s programming model when threads do independent work, it quickly breaks down in the presence of collective operations.

Unlike most instructions, which are executed by a single thread, collective operations must be executed collaboratively by a group of threads. For example, on line 14 of Figure 1, we perform a Tensor Core operation [43], which is only meaningful when invoked by a collection of 32 threads—a *warp*—acting together. For the call, each thread sets up its portion of the operands, and the operation is performed *once* for the entire warp, with the results scattered across participating threads. In this example, the first 32 threads in a block are tasked with this operation; programmers mentally group these 32 threads into a collective unit, and then reflect that grouping in the program via the `if`-statement on line 2. If the condition on line 2 were instead `tid >= 0 && tid < 30`, making fewer than 32 threads reach line 14, the result would be undefined. To make matters worse, the restriction is not only on the number of threads executing the operation, but also on their alignment. For threads to form a warp, the starting ID of the group must be aligned to a multiple of 32. So, if the condition on line 2 were instead `tid >= 1 && tid < 33`, the result would still be invalid, even though 32 threads would execute it.

Collective operations make reasoning about CUDA programs challenging because they force programmers to track the *convergence behavior* of threads. Specifically, programmers must reason both about how many threads reach each in the program and how those threads are arranged, accounting for alignment. This difficulty is further amplified by two factors. First, programs may have multiple points of convergence, requiring programmers to mentally track the relative ID of a thread within a logical group as that group evolves over the course of a program’s execution. Second, threads may be participating in multiple levels of convergence within the same program. In our example, we only considered the warp-level Tensor Core operation, but there are other collective operations that require convergence at different granularities like *warpgroup*-level Tensor Core operations—which must be issued collectively by four warps—or the block-level `__syncthreads()` primitive—which must be executed by all threads within a block.

Reasoning about collective operations is already error prone within the context of a single function, but becomes even more difficult when reasoning across functions. A callee may assume a certain configuration of threads and may structure its computation around that assumption. However, such assumptions are not visible in the callee’s function interface, which only exposes the input and output types. To invoke the function correctly, users must read its documentation—or worse, its implementation—to understand its assumptions, breaking modular reasoning.

In Prism, we materialize the programmer’s mental grouping of the compute hierarchy explicitly in the program’s source. Consider the example in Figure 2, which shows an equivalent rewrite of the CUDA program in Figure 1 using Prism. The call to the `mma` is valid only because it is executed within a **group** of 32 threads, made explicit on line 2. This intrinsic exposes its invariant via its function signature, which we will discuss in detail in Section 3.6. Since Prism’s **group** construct enforces both the size and alignment of participating threads, the validity of the `mma` operation is guaranteed by construction.

However, it is not sufficient to only consider the compute hierarchy when reasoning about convergence; we must consider the memory hierarchy as well. Whenever threads see different data, they can use it to induce divergent behavior. We’ve already seen an example of this: the `tid` variable in Figure 1. In general, the possibility of divergence lurks anywhere threads branch on data.

2.2 Data and the Memory Hierarchy

Data on the GPU is organized into a *memory hierarchy*, mirroring the compute hierarchy. All threads in a grid can read and write from *global memory*, where operands reside at the start of a computation and where results are eventually written. Each block has access to a limited amount of *shared memory*, a programmer-managed scratchpad typically used to stage repeatedly-used data. Finally, each thread maintains its own state in *registers* and *local memory*. These are private to each thread, while shared and global memory are accessible by multiple threads.

Thread-Local Data. Registers and local memory are owned by individual threads, so variables with the same name can have different values on different threads at run time. In general, CUDA offers no support for dealing with the divergence that arises as a consequence.

A key challenge for Prism is managing this divergence, ensuring that all threads organized by other language constructs like **group** remain logically unified, even in the face of data-dependent control flow. To do this, Prism tracks the *frequency* with which values vary in space. The `@` syntax attached to each declaration denotes this frequency. For example, `thread[1]` variables can vary for every thread, while `block[1]` variables can vary for every block, but not for threads within that block. Intuitively, the `@` construct “colors” a variable across threads, and any threads that share a color are required to agree on that variable’s value. Figure 3 illustrates valid and invalid colorings of a variable.

Using this information, Prism enforces rules for reading from and writing to variables. So, for example, if we introduce a condition based on `tid` inside the **group**’s scope in Figure 2, Prism will

reject that program; code that diverges at low frequency cannot read from variables that change at higher frequency. We explain the rules of Prism programs in detail in Section 3.

Thread-Shared Data. Shared and global memory, on the other hand, is not replicated per-thread; instead, all threads in a block have the same view into shared memory, while all threads on the grid have the same view into global memory.

CUDA does not explicitly model shared and global memory spaces, nor does it track whether allocations in a given memory space exceed device limits. Prism, on the other hand, distinguishes these spaces and restricts shared memory to static allocations, throwing an error if an allocation exceeds device limits. We will discuss these aspects in detail in Section 3.4. For now, however, we focus on how views of memory converge and diverge in tandem with the compute hierarchy.

As an example, when launching a kernel on the GPU, a global memory pointer initially belongs to the grid because every thread sees the same pointer value at the start of the computation. To write to that memory, each thread computes an offset from the original base address. In doing so, the pointer effectively diverges across threads so each can write independently. After these writes, the view on the memory must reconverge, ensuring that all threads have completed their updates before it can return to its original logical owner.

Let us reconsider the Tensor Core example from Figure 1, this time initializing the operands of the Tensor Core operation from pointers into global memory. In Figure 4, **A** and **B** are now populated from **a_mem** and **b_mem**. After the Tensor Core operation completes, the result is written back to **c_mem**, also in global memory. This variable **c_mem** is logically accessed from two different levels of the compute hierarchy. At the beginning, each thread in a group of 32 sees the same value for **c_mem**. Then, they each locate an offset within **c_mem** that they write to (lines 20–23). To restore **c_mem** back to its 32 thread-level ownership, all 32 threads must synchronize (line 25) to ensure all per-thread writes have completed. This is similar to the requirement we saw documented in CUB in Section 1.

Similarly to the compute hierarchy, CUDA programmers are responsible for tracking the logical owner of a view of memory as it evolves over the course of a program, and for ensuring that appropriate synchronization occurs.

Prism, by contrast, makes the evolution of memory’s view explicit in the syntax and automatically inserts the synchronization required to restore that view to its original owner. In Figure 5, we show how memory is lowered through the compute hierarchy for the same Tensor Core operation from Figure 4.

The lowering of **c_mem** is required in this program as Prism only permits writes from the view of a single thread. To lower **c_mem**, we use Prism’s **partition** operation on line 28. The operation takes a

```

1 int x = threadIdx.x;
2 if (x >= 0 && x < 32){
3     float A[4], B[2];
4     float C[4] = { 0 };
5
6     // k is a_mem's stride.
7     A[0] = a_mem[(x/4)*k+(x%4)];
8     A[1] = a_mem[(x/4)*k+(x%4)+4];
9     A[2] = a_mem[(x/4+8)*k+(x%4)];
10    A[3] = a_mem[(x/4+8)*k+(x%4)+4];
11
12    // n is b_mem's stride.
13    B[0] = b_mem[(x%4)*n+(x/4)];
14    B[1] = b_mem[(x%4+4)*n+(x/4)];
15
16    asm("mma.sync...");
17
18    // Write back into c_mem.
19    // n is c_mem's stride.
20    c_mem[(x/4)*n+2*(x%4)] = C[0];
21    c_mem[(x/4)*n+2*(x%4)+1] = C[1];
22    c_mem[(x/4+8)*n+2*(x%4)] = C[2];
23    c_mem[(x/4+8)*n+2*(x%4)+1]=C[3];
24
25    __syncwarp();
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
709
710
711
712
713
714
715
716
717
718
719
719
720
721
722
723
724
725
726
727
728
729
729
730
731
732
733
734
735
736
737
738
739
739
740
741
742
743
744
745
746
747
748
749
749
750
751
752
753
754
755
756
757
758
759
759
760
761
762
763
764
765
766
767
768
769
769
770
771
772
773
774
775
776
777
778
779
779
780
781
782
783
784
785
786
787
787
788
789
789
790
791
792
793
794
795
796
797
797
798
799
799
800
801
802
803
804
805
806
807
808
809
809
810
811
812
813
814
815
816
817
817
818
819
819
820
821
822
823
824
825
825
826
827
827
828
829
829
830
831
832
833
833
834
835
835
836
837
837
838
839
839
840
841
841
842
843
843
844
845
845
846
847
847
848
849
849
850
851
851
852
853
853
854
855
855
856
857
857
858
859
859
860
861
861
862
863
863
864
865
865
866
867
867
868
869
869
870
871
871
872
873
873
874
875
875
876
877
877
878
879
879
880
881
881
882
883
883
884
885
885
886
887
887
888
889
889
890
891
891
892
893
893
894
895
895
896
897
897
898
899
899
900
901
901
902
903
903
904
905
905
906
907
907
908
909
909
910
911
911
912
913
913
914
915
915
916
917
917
918
919
919
920
921
921
922
923
923
924
925
925
926
927
927
928
929
929
930
931
931
932
933
933
934
935
935
936
937
937
938
939
939
940
941
941
942
943
943
944
945
945
946
947
947
948
949
949
950
951
951
952
953
953
954
955
955
956
957
957
958
959
959
960
961
961
962
963
963
964
965
965
966
967
967
968
969
969
970
971
971
972
973
973
974
975
975
976
977
977
978
979
979
980
981
981
982
983
983
984
985
985
986
987
987
988
989
989
990
991
991
992
993
993
994
995
995
996
997
997
998
999
999
1000
1000
1001
1001
1002
1002
1003
1003
1004
1004
1005
1005
1006
1006
1007
1007
1008
1008
1009
1009
1010
1010
1011
1011
1012
1012
1013
1013
1014
1014
1015
1015
1016
1016
1017
1017
1018
1018
1019
1019
1020
1020
1021
1021
1022
1022
1023
1023
1024
1024
1025
1025
1026
1026
1027
1027
1028
1028
1029
1029
1030
1030
1031
1031
1032
1032
1033
1033
1034
1034
1035
1035
1036
1036
1037
1037
1038
1038
1039
1039
1040
1040
1041
1041
1042
1042
1043
1043
1044
1044
1045
1045
1046
1046
1047
1047
1048
1048
1049
1049
1050
1050
1051
1051
1052
1052
1053
1053
1054
1054
1055
1055
1056
1056
1057
1057
1058
1058
1059
1059
1060
1060
1061
1061
1062
1062
1063
1063
1064
1064
1065
1065
1066
1066
1067
1067
1068
1068
1069
1069
1070
1070
1071
1071
1072
1072
1073
1073
1074
1074
1075
1075
1076
1076
1077
1077
1078
1078
1079
1079
1080
1080
1081
1081
1082
1082
1083
1083
1084
1084
1085
1085
1086
1086
1087
1087
1088
1088
1089
1089
1090
1090
1091
1091
1092
1092
1093
1093
1094
1094
1095
1095
1096
1096
1097
1097
1098
1098
1099
1099
1100
1100
1101
1101
1102
1102
1103
1103
1104
1104
1105
1105
1106
1106
1107
1107
1108
1108
1109
1109
1110
1110
1111
1111
1112
1112
1113
1113
1114
1114
1115
1115
1116
1116
1117
1117
1118
1118
1119
1119
1120
1120
1121
1121
1122
1122
1123
1123
1124
1124
1125
1125
1126
1126
1127
1127
1128
1128
1129
1129
1130
1130
1131
1131
1132
1132
1133
1133
1134
1134
1135
1135
1136
1136
1137
1137
1138
1138
1139
1139
1140
1140
1141
1141
1142
1142
1143
1143
1144
1144
1145
1145
1146
1146
1147
1147
1148
1148
1149
1149
1150
1150
1151
1151
1152
1152
1153
1153
1154
1154
1155
1155
1156
1156
1157
1157
1158
1158
1159
1159
1160
1160
1161
1161
1162
1162
1163
1163
1164
1164
1165
1165
1166
1166
1167
1167
1168
1168
1169
1169
1170
1170
1171
1171
1172
1172
1173
1173
1174
1174
1175
1175
1176
1176
1177
1177
1178
1178
1179
1179
1180
1180
1181
1181
1182
1182
1183
1183
1184
1184
1185
1185
1186
1186
1187
1187
1188
1188
1189
1189
1190
1190
1191
1191
1192
1192
1193
1193
1194
1194
1195
1195
1196
1196
1197
1197
1198
1198
1199
1199
1200
1200
1201
1201
1202
1202
1203
1203
1204
1204
1205
1205
1206
1206
1207
1207
1208
1208
1209
1209
1210
1210
1211
1211
1212
1212
1213
1213
1214
1214
1215
1215
1216
1216
1217
1217
1218
1218
1219
1219
1220
1220
1221
1221
1222
1222
1223
1223
1224
1224
1225
1225
1226
1226
1227
1227
1228
1228
1229
1229
1230
1230
1231
1231
1232
1232
1233
1233
1234
1234
1235
1235
1236
1236
1237
1237
1238
1238
1239
1239
1240
1240
1241
1241
1242
1242
1243
1243
1244
1244
1245
1245
1246
1246
1247
1247
1248
1248
1249
1249
1250
1250
1251
1251
1252
1252
1253
1253
1254
1254
1255
1255
1256
1256
1257
1257
1258
1258
1259
1259
1260
1260
1261
1261
1262
1262
1263
1263
1264
1264
1265
1265
1266
1266
1267
1267
1268
1268
1269
1269
1270
1270
1271
1271
1272
1272
1273
1273
1274
1274
1275
1275
1276
1276
1277
1277
1278
1278
1279
1279
1280
1280
1281
1281
1282
1282
1283
1283
1284
1284
1285
1285
1286
1286
1287
1287
1288
1288
1289
1289
1290
1290
1291
1291
1292
1292
1293
1293
1294
1294
1295
1295
1296
1296
1297
1297
1298
1298
1299
1299
1300
1300
1301
1301
1302
1302
1303
1303
1304
1304
1305
1305
1306
1306
1307
1307
1308
1308
1309
1309
1310
1310
1311
1311
1312
1312
1313
1313
1314
1314
1315
1315
1316
1316
1317
1317
1318
1318
1319
1319
1320
1320
1321
1321
1322
1322
1323
1323
1324
1324
1325
1325
1326
1326
1327
1327
1328
1328
1329
1329
1330
1330
1331
1331
1332
1332
1333
1333
1334
1334
1335
1335
1336
1336
1337
1337
1338
1338
1339
1339
1340
1340
1341
1341
1342
1342
1343
1343
1344
1344
1345
1345
1346
1346
1347
1347
1348
1348
1349
1349
1350
1350
1351
1351
1352
1352
1353
1353
1354
1354
1355
1355
1356
1356
1357
1357
1358
1358
1359
1359
1360
1360
1361
1361
1362
1362
1363
1363
1364
1364
1365
1365
1366
1366
1367
1367
1368
1368
1369
1369
1370
1370
1371
1371
1372
1372
1373
1373
1374
1374
1375
1375
1376
1376
1377
1377
1378
1378
1379
1379
1380
1380
1381
1381
1382
1382
1383
1383
1384
1384
1385
1385
1386
1386
1387
1387
1388
1388
1389
1389
1390
1390
1391
1391
1392
1392
1393
1393
1394
1394
1395
1395
1396
1396
1397
1397
1398
1398
1399
1399
1400
1400
1401
1401
1402
1402
1403
1403
1404
1404
1405
1405
1406
1406
1407
1407
1408
1408
1409
1409
1410
1410
1411
1411
1412
1412
1413
1413
1414
1414
1415
1415
1416
1416
1417
1417
1418
1418
1419
1419
1420
1420
1421
1421
1422
1422
1423
1423
1424
1424
1425
1425
1426
1426
1427
1427
1428
1428
1429
1429
1430
1430
1431
1431
1432
1432
1433
1433
1434
1434
1435
1435
1436
1436
1437
1437
1438
1438
1439
1439
1440
1440
1441
1441
1442
1442
1443
1443
1444
1444
1445
1445
1446
1446
1447
1447
1448
1448
1449
1449
1450
1450
1451
1451
1452
1452
1453
1453
1454
1454
1455
1455
1456
1456
1457
1457
1458
1458
1459
1459
1460
1460
1461
1461
1462
1462
1463
1463
1464
1464
1465
1465
1466
1466
1467
1467
1468
1468
1469
1469
1470
1470
1471
1471
1472
1472
1473
1473
1474
1474
1475
1475
1476
1476
1477
1477
1478
1478
1479
1479
1480
1480
1481
1481
1482
1482
1483
1483
1484
1484
1485
1485
1486
1486
1487
1487
1488
1488
1489
1489
1490
1490
1491
1491
1492
1492
1493
1493
1494
1494
1495
1495
1496
1496
1497
1497
1498
1498
1499
1499
1500
1500
1501
1501
1502
1502
1503
1503
1504
1504
1505
1505
1506
1506
1507
1507
1508
1508
1509
1509
1510
1510
1511
1511
1512
1512
1513
1513
1514
1514
1515
1515
1516
1516
1517
1517
1518
1518
1519
1519
1520
1520
1521
1521
1522
1522
1523
1523
1524
1524
1525
1525
1526
1526
1527
1527
1528
1528
1529
1529
1530
1530
1531
1531
1532
1532
1533
1533
1534
1534
1535
1535
1536
1536
1537
1537
1538
1538
1539
1539
1540
1540
1541
1541
1542
1542
1543
1543
1544
1544
1545
1545
1546
1546
1547
1547
1548
1548
1549
1549
1550
1550
1551
1551
1552
1552
1553
1553
1554
1554
1555
1555
1556
1556
1557
1557
1558
1558
1559
1559
1560
1560
1561
1561
1562
1562
1563
1563
1564
1564
1565
1565
1566
1566
1567
1567
1568
1568
1569
1569
1570
1570
1571
1571
1572
1572
1573
1573
1574
1574
1575
1575
1576
1576
1577
1577
1578
1578
1579
1579
1580
1580
1581
1581
1582
1582
1583
1583
1584
1584
1585
1585
1586
1586
1587
1587
1588
1588
1589
1589
1590
1590
1591
1591
1592
1592
1593
1593
1594
1594
1595
1595
1596
1596
1597
1597
1598
1598
1599
1599
1600
1600
1601
1601
1602
1602
1603
1603
1604
1604
1605
1605
1606
1606
1607
1607
1608
1608
1609
1609
1610
1610
1611
1611
1612
1612
1613
1613
1614
1614
1615
1615
1616
1616
1617
1617
1618
1618
1619
1619
1620
1620
1621
1621
1622
1622
1623
1623
1624
1624
1625
1625
1626
1626
1627
1627
1628
1628
1629
1629
1630
1630
1631
1631
1632
1632
1633
1633
1634
1634
1635
1635
1636
1636
1637
1637
1638
1638
1639
1639
1640
1640
1641
1641
1642
1642
1643
1643
1644
1644
1645
1645
1646
1646
1647
1647
1648
1648
1649
1649
1650
1650
1651
1651
1652
1652
1653
1653
1654
1654
1655
1655
1656
1656
1657
1657
1658
1658
1659
1659
1660
1660
1661
1661
1662
1662
1663
1663
1664
1664
1665
1665
1666
1666
1667
1667
1668
1668
1669
1669
1670
1670
1671
1671
1672
1672
1673
1673
1674
1674
1675
1675
1676
1676
1677
16
```

```

1  @prism("global")
2  # Top-level perspective bounds and shared memory usage.
3  @requires(grid[1], block[1], thread[32], smem=1280)
4  def mmaTF32NaiveKernel(A: ptr(const(float)) @ grid[1],
5                           B: ptr(const(float)) @ grid[1],
6                           C: ptr(float) @ grid[1],
7                           M : int @ grid[1],
8                           N : int @ grid[1],
9                           K : int @ grid[1]):
10
11    # Starts out with grid[1] perspective.
12    with group(grid[1]):
13      # @ grid[1] inferred from current perspective
14      # Each block computes an 16 x 8 tile
15      num_blocks_n : const(int) = (N + 8 - 1) / 8
16
17      # id() function returns the block id
18      # inferred from @ block[1].
19      blk_row : const(int) @ block[1] = (id() / num_blocks_n) * 16
20      blk_col : const(int) @ block[1] = (id() % num_blocks_n) * 8
21
22      # Give each block an offset into C
23      offset = lambda x: blk_row * N + blk_col + x
24      with partition(C, p=block[1], f=offset) as C_blk:
25        with group(block[1]):
26          # SHMEM declarations are only allowed
27          # with a block[1] perspective.
28          A_smem : shared(float[16 * 8]) @ block[1]
29          B_smem : shared(float[8 * 8]) @ block[1]
30          C_smem : shared(float[16 * 8]) @ block[1]
31          # Now, id() returns the thread id
32          idx = lambda x: x * 4
33          # To write to C_smem, drop to thread[1] perspective
34          with partition(C_smem, p=thread[1], f=idx) as C_th:
35            for i in range(0, 4, 1):
36              with group(thread[1]):
37                C_th[i] = 0
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
for i in range(0, K_tiles, 1):
  # --- Sync point --- (backedge from for loop)
  for j in range(0, 4, 1):
    global_row : int @ thread[1] = blk_row + row
    global_col : int @ thread[1] = i * 8 + col
    with partition(A_smem, p=thread[1], f=...) as A_th:
      with group(thread[1]):
        A_thrd[0] = A[global_row * K + global_col]

  # --- Sync point --- (backedge from for loop)
  for j in range(0, 2, 1):
    # Similar to write into A_smem ...
    with partition(B_smem, p=thread[1], f=...) as B_th:
      with group(thread[1]):
        B_th[0] = B[global_row_b * N + global_col_b]

    # Give each warp an offset into C_smem.
    # --- Sync point --- (backedge from for loop)
    with claim(C_smem, p=thread[32]) as Cs_warp:
      match split(thread): # Masks off other threads
        case 32:
          # Call function that performs a
          # Tensor Core instruction.
          simple_mma(A_smem, B_smem, Cs_warp)
    # --- Sync point ---

  for j in range(0, 4, 1):
    flat_idx_c : int @ thread[1] = id() * 4 + j
    row_c : int @ thread[1] = flat_idx_c / MMA_K
    col_c : int @ thread[1] = flat_idx_c % MMA_K
    idx = lambda x: row_c * N + col_c + x
    with partition(C_blk, p=thread[1], f=idx) as C_th:
      with group(thread[1]):
        C_th[0] = C_smem[row_c * MMA_N + col_c]

return

```

Fig. 6. Naive tensor float 32 matrix multiplication in Prism (full program can be found in Appendix B.1).

pointer to partition, **c_mem**, and lowers it to a single thread, assigning it a new name, **c_thrd**; Prism will not allow use of the old variable **c_mem** within the **partition**'s scope. The **partition** also takes an indexing function, and each time **c_thrd** is accessed, this function is implicitly applied. Once the **partition**'s scope ends, Prism inserts a synchronization barrier before the next use of the original variable, so that all per-thread writes have completed. In this way, at the end of the **partition**, the original variable represents a convergent view of the data once again.

3 The Prism Language

Prism is an imperative, low-level language designed at a level of abstraction comparable to that of CUDA. Unlike CUDA, however, Prism's syntax materializes the mapping of computations onto the compute and memory hierarchy explicitly in the program source. Using this information, Prism enforces that programs only execute collective operations with sufficient resources.

To guide our discussion about Prism's design, we use the program in Figure 6 as a running example. It computes a matrix multiplication between two float arrays, **A** and **B**, to produce an output matrix **C**. In this program, each block computes an independent 16×8 tile of the output. To do so, blocks first locate the tile index assigned to them (line 19-20). Next, threads in each block load corresponding rows and columns from **A** and **B** (line 41-53) into shared memory. Finally, the program invokes a warp-level, Tensor Core instruction to compute the output (line 62), requiring threads in a warp to converge. We encapsulate this Tensor Core instruction in a function, demonstrating how function composition works in Prism. This function is the same as the one in Figure 5.

3.1 Levels

Prism models the machine's compute hierarchy through *levels*. There are three levels in Prism—**grid**, **block** and **thread**—which are organized as expected: a **grid** consists of multiple **blocks**, each of which consist of multiple **threads**. Levels are ordered, with **thread < block < grid**.

There are two key differences between Prism's levels and those in CUDA. First, Prism does not model CUDA's three-dimensional grid or block structure. Second, there are two commonly-used "levels", warp and warpgroup, absent from our hierarchy. On the hardware, the units of each level are arranged in a linear order, and the three-dimensional structures of CUDA are simply interpretations of this ordering, not distinct hardware resources. Similarly, warps and warpgroups are organizational constructs defined in terms of existing levels. Namely, a warp is a group of 32 threads whose first thread ID is aligned to 32. A warpgroup, which was introduced with the release of the Hopper architecture [8], consists of 4 consecutive warps.

Rather than baking these interpretations into Prism by adding new dimensions and levels, we let users express multi-dimensional structures and define groupings of custom sizes.

3.2 Perspectives

Perspectives are the central concept of Prism, representing the view of the hierarchy from which a given statement is defining the machine's behavior. Perspectives allow Prism to determine which compute resources the programmer is controlling at every point in the program, whether they are available in the program's context, and if those resources are sufficient for a given operation.

A perspective is a level—**grid**, **block**, or **thread**—paired with a static constant **n**, specifying the number of units at that level. For example, **thread[2]** denotes a perspective of two threads, **block[4]** denotes a perspective of four blocks, and so on. Perspectives also carry *alignment information*: a perspective of size **n** is aligned to **n**. In this way, a warp is simply a desugaring of **thread[32]**, and a warpgroup is a desugaring of **thread[128]**.

Finally, perspectives are partially ordered. We say that **level₂[n₂]** is *broader* than **level₁[n₁]**, or that **level₁[n₁]** is *narrower* than **level₂[n₂]**, if either:

- (1) **level₁** < **level₂**; or,
- (2) **level₁** = **level₂** and **n₁** divides **n₂**.

3.3 Perspectives on Code

Code is associated with a set of perspectives, called a *perspective bound*, which corresponds to the compute resources whose behavior it defines. At every point in the program, the perspective bound for that point indicates which layer of the hierarchy is being programmed, and how that layer can be destructed into narrower perspectives. For example, a line of code with a **{block[1], thread[4]}** perspective bound tells Prism that the current line of code is being programmed at the **block[1]** perspective and that the **block[1]** perspective can be destructed into a some number of **thread[4]** perspectives. As a shorthand, when we refer to a code's perspective, we mean the broadest perspective available in its perspective bound (in this example, **block[1]**).

Functions begin with a top-level perspective bound. In Figure 6, the perspective bound is defined on line 2, using the notation **@requires(grid[n₁], block[n₂], thread[n₃])**. Programmers can then shape the program's current perspective using two constructs: **group** and **split**.

Group. The **group** construct lets programmers shift from a broader perspective to some number of narrower perspectives contained in it. Operationally, the **group** construct does this by replicating code written from the narrower perspective across the broader one. In effect, **group** forks many copies of a narrower perspective.

Let's consider this in context of our example. In Figure 6, execution begins at **grid[1]** on line 12. At that point, a programmer controls the whole grid's behavior. To produce different output tiles, the programmer shifts their perspective to **block[1]** on line 25. The code within the **group(block[1])** defines the behavior of a single block, and is replicated across all blocks in the grid.

```

1 # Example 1
2 with group(thread[2]):
3     # Illegal because block > thread.
4     with group(block[1]):
5         pass
6 # Example 2
7 with group(block[6]):
8     # Illegal because 6 % 5 != 0.
9     with group([block[5]]):
10        pass

```

Fig. 7. Illegal uses of **group**.

Not all uses of **group** are valid: the examples in Figure 7 have no meaning on the hardware. On line 4, the program tries to broaden its perspective to **block[1]** from **thread[2]**, which is illegal. On the other hand, in the second example, the program tries to narrow its perspective from **block[6]** to **block[5]**. While 5 is indeed less than 6, Prism cannot evenly replicate **block[5]** across a **block[6]** perspective, and so rejects this program. Recall, from Section 3.2, that whether one perspective is narrower than another is dependent on divisibility, not just size.

To eliminate such cases, Prism only allows an invocation of **group(level[n])** if the current perspective bound contains a perspective broader than **level[n]**. Once **group(level[n])** is invoked, it modifies the current perspective bound in two ways. First, it removes all perspectives broader than **level[n]** from it. Second, it sets the broadest perspective within the **group** to be **level[n]**.

Split. Unlike **group**, which is used for replication into equally-sized, narrower perspectives, **split** is used for sharding the current perspective unequally. For example, Figure 8 shows a **split** from **thread[4]** into one branch with **thread[2]** perspective and two with **thread[1]**. The three arms of the **split** execute independently in parallel,¹ as a form of unordered composition. When **split(level)** is invoked, the perspectives of its branches diverge. At the end of the **split**, they reconverge and continue execution with the original perspective.

Use of **split** is necessary to write warp-specialized [4] code, a programming pattern used in high-performance kernels. Another important use of **split**—masking off threads—can be found in our running example. Line 58 in Figure 6 shows a **split** that requests the first warp in the block, narrowing from **block[1]** into a single **thread[32]**. This warp is later used to execute a Tensor Core operation.

Because **split** corresponds to unordered composition, it must provide each of its branches their requested perspectives simultaneously. Prism thus checks that the sum of the perspectives requested by all branches of the **split** can be satisfied. For example, the program in Figure 9 does not type check. Finally, because perspectives enforce alignment, every branch of the **split** must also be aligned; not all **splits** whose sizes are at most the available units are valid. Figure 10 shows an example violating this constraint: the second branch of the split is not aligned to 2.

Once **split** is invoked, for each branch that requests **n** units, all perspectives broader than **level[n]** are removed from its perspective bound, and the available units for **level** are set to **n**.

3.4 Perspectives on Data

Section 3.2 described how programmers can control different layers of the hierarchy by changing their perspectives on code through **group** and **split**. To ensure these operations remain meaningful, Prism must ensure that threads inside a perspective remain logically grouped, even when they encounter **for**, **while**, and **if** statements. As we saw in Section 2.2, making this guarantee requires Prism to track how data varies across threads.

3.4.1 Thread-Local Data. In Prism, each local variable has a perspective, which indicates the frequency at which its values change in space. This frequency remains constant for the duration of a program, and it tells Prism that a **level[n]** variable is always indistinguishable to threads within that perspective. For example, **blk_row @ block[1]** on line 19 in Figure 6 has the same value across all threads in a block.

```
1 with group(thread[4]):
2   match split(thread):
3     case 2:
4       ...
5     case 1:
6       ...
7     case 1:
8       ...
```

Fig. 8. Example **split**.

```
1 with group(thread[4]):
2   match split(thread):
3     case 4:
4       ...
5     case 1:
6       ...
```

Fig. 9. Illegal split exceeds the available perspectives.

```
1 with group(thread[3]):
2   match split(thread):
3     case 1:
4       ...
5     case 2:
6       ...
```

Fig. 10. Illegal split violates alignment.

¹Despite the use of the **match** syntax, *all* branches of the **split** execute.

Programmers specify the perspective that each variable lives at in its declaration. A variable `v` of type `int` is declared at `thread[1]` perspective using the syntax `v : int @ thread[1]`.² To enforce this frequency invariant, Prism restricts reads from and writes to variables based on their perspectives. The rule can be summarized as follows: Prism allows programs to “read up” from broader perspectives and “write down” to narrower ones.

Read Up. Variables can only be read if their perspective is at least as broad as the current code perspective. Figure 11 gives an example of an illegal read that would violate this constraint. While `__syncthreads()` should always be safe in `block[1]`, branching on the variable `flag`—which may take different values across threads in a block—can cause only some of those threads to reach the `__syncthreads()`, violating its collective invariant.

```

1 flag : bool @ thread[1] = ...
2 with group(block[1]):
3   if (flag)
4     __syncthreads();

```

Fig. 11. Illegal read of `thread[1]` variable.

Write Down. Writes are dually constrained. Only values that live at broader perspectives can be written into variables that live at narrower ones. For example, a `block[1]` variable can be used to write into a `thread[1]` variable, but not vice versa. An example of an illegal write is shown in Figure 12, where writing from a `thread[1]` variable into a `block[1]` one would cause a deadlock.

```

1 x : bool @ thread[1] = ...
2 y : bool @ block[1] = ...
3 with group(block[1]):
4   y = x
5   if(y):
6     __syncthreads();

```

Fig. 12. Illegal write of `thread[1]` variable into a `block[1]` variable.

Together, the “read up” and “write down” rules ensure that information only flows from broader perspectives to narrower ones. In Figure 6, we can see the “read up” rule in action on lines 42 and 43. Meanwhile, lines 19 and 20 are instances of the “write down” rule.

3.4.2 Thread-Shared Data. Unlike thread-local data, which is literally replicated across threads and is backed by distinct physical storage, thread-shared data consists of pointers into shared and global memory that are visible to some collection of threads. As a result, the perspective such data inhabits can evolve as the program executes.

In Prism, there are three mechanisms for obtaining thread-shared data. The first is to directly allocate data residing in shared memory.³ The second and third are to obtain offsets into existing data by using the `partition` or `claim` operations. These constructs mirror `group` and `split`, and are used to temporarily narrow the perspective associated with a pointer.

Prism simplifies shared memory allocations by requiring them to have a static size. The `@requires` annotation specifies the amount of shared memory a function expects to have. Prism uses this information to ensure that a function’s allocations do not exceed its declared limit, and checks whether there is enough shared memory available at its call sites. The function in Figure 6 declares the amount of shared memory it will require on 2. We use standard techniques [34] to statically bound memory usage with Prism’s type system.

Additionally, since shared memory is only visible to threads within the same block, allocations to it are only permitted at `block[1]`. An example of such an allocation is shown on lines 28–30 of Figure 6. During compilation, Prism will automatically handle the necessary pointer arithmetic to assign each allocation an appropriate offset within the shared memory space.

Partition. The `partition` operation plays the same role for memory that `group` plays for code. It is used to refine the perspective of a pointer by computing offsets for each narrowed perspective. Concretely, the `partition` operation takes a pointer variable `x`, an indexing function `f`, and produces

²If not explicitly annotated, Prism infers a variable’s perspective to be the perspective of the code where it was declared.

³Prism assumes that all global memory allocations have been made before launching the CUDA kernel, as is typically the case with CUDA programs.

a new variable **y** at a narrower perspective **level[n]** to be used inside the **partition**. After the partition ends, the threads participating in the **partition** must synchronize to restore the original pointer **x** to its perspective.⁴

Inside the scope of the partition, the original variable **x** cannot be used, and the pointer can only be accessed using **y**. Each use of **y** applies the indexing function **f** to compute the true offset of the access. For example, the use of **partition** on line 34 in Figure 6 takes a pointer **C_smem** at **block[1]**, gives it a new name **C_th**, and distributes it across **thread[1]** perspectives by transforming each occurrence of **C_th[i]** in the body into **C_smem[i*4]**.

At this point, it is necessary to consider how different indexing functions affect the possibility of data races. If the indexing function is injective, each narrower perspective receives a distinct offset into the underlying array, and the resulting partition is free of data races. When the indexing function is not injective, multiple threads may race on the same location, introducing a potential data race.

Preventing data races is not one of Prism's goals. In Prism, data races *can* occur within a **partition**; however, since the data is eventually synchronized before it is reused, the last writer wins. Out-of-bounds accesses are considered undefined behavior. Prior work, in particular Descend [35], describes a type system for restricting data-races and out-of-bounds accesses, and we believe a similar approach can be combined with Prism's. Prism instead focuses on the interaction between the compute and memory hierarchies and on reasoning about them simultaneously to ensure that operations are executed only with sufficient compute resources, a guarantee that Descend cannot provide.

Claim. The **claim** operation lets programmers narrow a pointer's perspective by giving it to only one collection of threads with that narrower perspective. For example, line 57 of Figure 6 shows an example of a **claim**, where the pointer **C_smem @ block[1]** is only available to a single warp in the block and narrowed to **Cs_warp @ thread[32]** to call a Tensor Core operation.

A **claim** takes the original pointer **x**, the target perspective to narrow it to, and a new name **y** to assign to the narrowed memory. In all **splits** within the **claim**, the new pointer **y** is accessible only within one branch; sibling branches are not permitted to read or write from this memory. As with **partition**, the original variable **x** is not accessible within the **claim**.

3.5 The **id()** Function

As opposed to exposing users to special hardware variables like **blockIdx.x**, and **threadIdx.x**, Prism provides an **id()** function instead. The **id** function returns the relative index of a narrow perspective within a broader perspective. That is, the interpretation of the **id** function depends on both the perspective of the variable it is being written to and the perspective of the code invoking it. In Figure 6, we use a call to **id** with **grid[1]** perspective on lines 19 and 20 to locate the tile that each block is in charge of computing. On line 32, however, a write of **id** into a **thread[1]** variable at **block[1]** perspective will return the relative ID of the thread within the block, not in the grid.

3.6 Collective Operations

There are two types of collective operations in Prism.

Function Calls. Each function carries a *perspective signature*, which consists of its top-level perspective bound, shared memory usage, and its arguments' perspectives. An example of such a signature can be seen in lines 3-9 in Figure 6. At call sites, Prism ensures that the callee's perspective signature can be satisfied by the caller. Since the perspective bound describes a minimum requirement, functions can be called with a broader code perspective than necessary, accounting for alignment. When checking

⁴Strictly speaking, this synchronization is only required at the next point the memory is to be *re-used*. Prism has a mechanism for optimizing the placement of these synchronization points, which is described in Section 5.

function arguments at call sites, Prism distinguishes between primitive data types and pointers. For primitive data types, like `int` or `bool`, arguments can have broader perspectives than specified in the function signature. On the other hand, pointers that live at broader perspectives can only be passed if the pointer is marked as `const`, indicating that it will only be used for reading. Otherwise, we require that the pointer’s perspective exactly match that of the function’s perspective signature.

Intrinsics. The compiler provides a pre-defined set of collective intrinsics—like the Tensor Core instruction discussed in Section 2.1—each of which declares its perspective signature. Programmers may also add to this set using `unsafe`. From within `unsafe` code, programmers can inline assembly instructions and wrap them in a Prism function, specifying its perspective signature. Prism checks these call sites like those of any other function.

3.7 Asynchrony

Asynchronous data movement works similarly to other memory operations. Users can mark storage as asynchronous with the `async` construct. As with `partition` and `claim`, `async` hides the old variable and exposes a fresh one that is only accessible within the `async` statement. Inside, Prism ensures the new variable is only used by `async` data-movement intrinsics.

Prism includes two such intrinsics: bulk [16] and non-bulk [19] asynchronous data-movement instructions. We show an example of the former in Figure 13. As with the data operations in Section 3.4, Prism inserts the necessary synchronization before the program’s next use of the original variable, ensuring that the asynchronous transfer has completed.

```

1 with async(old_name) as new_name:
2   match split(thread):
3     case 1:
4       cp.async(smem, new_name, 16);

```

Fig. 13. Example of an `cp.async` instruction in Prism.

4 Formalization in Bundl

Having introduced the full Prism language, we now describe Bundl, a core calculus that formalizes its most fundamental aspects by statically tracking perspectives on code and data. We use Bundl to argue that well-typed Prism programs are not only type-safe, but will also never execute operations for which they lack the correct perspective.

In this section, we describe Bundl’s type system and operational semantics—in particular how it manages compute and data perspectives—and build up to a formal proof of type-and-perspective safety. We instrument Bundl’s operational semantics with runtime perspective enforcement: its rules will get stuck if they encounter an operation for which they have the wrong perspective. This runtime enforcement means that our safety theorem guarantees that dynamically-realized perspectives match the ones inferred by the type system.

4.1 Bundl Type System

The core idea in Bundl is to track, at the type level, the program’s perspective on code and data. To achieve this, we borrow techniques from the literature on dependency tracking [1]. In particular, the code perspective is tracked on the typing judgment, which has the form $\Gamma \vdash^\pi e : \tau$ for expressions and $\Gamma \vdash^\pi s$ for statements. The π over the \vdash is the code perspective on e and s —and is comprised of a level and a size—the same structure as a perspective in Prism.

The typing context also tracks the perspective at which each variable lives; data can only be read from or written to a variable when its perspective is compatible with that of the code interacting with it. This requirement is made manifest in the `T-VAR` rule, found in Figure 14. Observe that the π in the variable rule must match exactly between data and code; principles like “read up” and “write down” are instead encoded directly in the rules for reading and writing, like `T-ARR-ACCESS`, which views the array being read with broader perspective than the current code perspective.

Figure 14 also shows other key rules, which fall into two main categories: those for managing perspectives on code and those for perspectives on data.

$\boxed{\Gamma \vdash^\pi e : \tau}$	<i>(Expression typing)</i>
$\frac{x : \pi \in \Gamma}{\Gamma \vdash^\pi x : \tau} \text{ T-VAR}$	$\frac{\Gamma \vdash^{\pi'} e_1 : \tau[]^l \quad \Gamma \vdash^\pi e_2 : \mathbf{int} \quad \pi \leq \pi'}{\Gamma \vdash^\pi e_1[e_2] : \tau} \text{ T-ARR-ACCESS}$
$\boxed{\Gamma \vdash^\pi s}$	<i>(Statement typing)</i>
	$\frac{\Gamma \vdash^{(h,n_1)} s_1 \quad \Gamma \vdash^{(h,n_2)} s_2 \quad n_1, n_2 \text{ align to } n}{\Gamma \vdash^{(h,n)} \mathbf{split}(n_1, n_2)\{s_1\}\{s_2\}} \text{ T-SPLIT}$
	where $n_1, n_2 \text{ align to } n ::= (n_1 + n_2 \leq n) \text{ and } (n_1 n) \text{ and } (n_2 n) \text{ and } (n_2 n_1 + n)$
$\frac{\Gamma \vdash^\pi s}{\Gamma \vdash^{q \cdot \pi} \mathbf{group} q s} \text{ T-GROUP}$	$\frac{\Gamma, y : \downarrow^\pi \tau[]^l \vdash^\pi s \quad l \neq \mathbf{Local}}{\Gamma, x : \pi \tau[]^l \vdash^\pi \mathbf{lower} x \text{ into } y \text{ in } s} \text{ T-LOWER}$
$\frac{\Gamma, y : (h, n/c) \tau[]^l \vdash^{(h,n)} s \quad c n \quad l \neq \mathbf{Local}}{\Gamma, x : (h,n) \tau[]^l \vdash^{(h,n)} \mathbf{partition} x \text{ into } y \text{ by } c \text{ in } s} \text{ T-PARTITION}$	$\frac{\Gamma \vdash^\pi \mathbf{destruct} \text{ in } s}{\Gamma \vdash^\pi s} \text{ T-DESTRUCT}$
	$\frac{\Gamma, y : (h, n') \tau[]^l \vdash^{(h,n')} s \quad n' \leq n \quad l \neq \mathbf{Local}}{\Gamma, x : (h,n) \tau[]^l \vdash^{(h,n)} \mathbf{claim} x \text{ into } y \text{ at } n' \text{ in } s} \text{ T-CLAIM}$

Fig. 14. Core typing rules of Bundl. The typing rules presented here are a simplified selection of the full rules, which can be found in Appendix A.2.

4.1.1 Managing Perspectives on Code. Bundl’s **group** statement directly corresponds to Prism’s, and is checked by the **T-GROUP** rule. Given some statement s that checks with perspective π , the statement **group** $q s$ will check with $q \cdot \pi$, enforcing Prism’s divisibility requirement.

The **split** statement, meanwhile, is checked by the **T-SPLIT** rule, and functions like a binary version of the n -ary **split** construct in Prism. It enforces the same divisibility requirements to ensure that the perspectives on code and data remain properly aligned, and then checks the two sub-statements s_1 and s_2 with the divided, narrower perspectives.

In Bundl, to better model the details of how perspectives shift down the GPU hierarchy, we introduce a third construct called **destruct** that makes explicit exactly where such shifts occur. In the corresponding rule, **T-DESTRUCT**, the \downarrow operation on π s “destructs” the perspective into many narrower perspectives at a lower level. This operation is defined as $\downarrow(\mathbf{Grid}, 1) = (\mathbf{Block}, B)$ and $\downarrow(\mathbf{Block}, 1) = (\mathbf{Thread}, T)$, where B and T are parameters to a particular instantiation of Bundl to describe the number of blocks per grid and threads per block.⁵

4.1.2 Managing Perspectives on Data. The mechanism for managing data perspective mirrors that of code perspective, with each operation for data corresponding to an operation for code.

The **partition** operation is analogous to **grouping** a code perspective. The typing rule for this operation, **T-PARTITION**, requires that the data perspective on x , the variable to be partitioned, is the same as the current perspective on code. After partitioning, a fresh variable y is introduced with a new perspective π/c . Within the **partition**, we disallow references to x and continue checking the body with the original π ; the **partition** has no effect on the code perspective.

Unlike **partition**, which divides up a piece of data equally among narrower perspectives, the **claim** operation views the claimed data with exactly one narrower perspective. Accordingly, Bundl needs to ensure that only one branch of a **split** operation, with the appropriate π , can refer to the claimed variable. To ensure that this is the case, the **T-CLAIM** rule links the data perspective of the variable to the compute perspective of the code claiming it by changing both at the same time.

⁵Bundl is abstracted over these B and T values, so instead of tracking perspective bounds the way that Prism does, it only tracks the top-level perspective described in Section 3.2.

$$\begin{array}{c}
\boxed{L,S,\Sigma,P \rightsquigarrow L',S',\Sigma',P'} \quad (Machine\ judgment) \\
\frac{L(t),S(b),\Sigma,t,b,0 \vdash^{(\texttt{Grid},1)} s \rightsquigarrow s' + \eta',\sigma',\Sigma' \quad P(t,b) = s}{L,S,\Sigma,P \rightsquigarrow L[t \mapsto \eta'],S[b \mapsto \sigma'],\Sigma',P[(t,b) \mapsto s']} \text{ S-PROGRAM} \\
\\
\boxed{\eta,\sigma,\Sigma,t,b,p \vdash^\pi s \rightsquigarrow s' + \eta',\sigma',\Sigma'} \quad (Thread\ judgment) \\
\\
\frac{p < n_1 \quad n_1,n_2 \text{ align to } n \quad \eta,\sigma,\Sigma,t,b,p \vdash^{(h,n_1)} s_1 \rightsquigarrow s'_1 + \eta',\sigma',\Sigma'}{\eta,\sigma,\Sigma,t,b,p \vdash^{(h,n)} \text{split}(n_1,n_2)\{s_1\}\{s_2\} \rightsquigarrow \text{split}(n_1,n_2)\{s'_1\}\{s_2\} + \eta',\sigma',\Sigma'} \text{ S-SPLIT-LEFT} \\
\\
\frac{p \geq n_1 \quad p < n_1+n_2 \quad n_1,n_2 \text{ align to } n \quad \eta,\sigma,\Sigma,t,b,p - n_1, \vdash^{(h,n_2)} s_2 \rightsquigarrow s'_2 + \eta',\sigma',\Sigma'}{\eta,\sigma,\Sigma,t,b,p \vdash^{(h,n)} \text{split}(n_1,n_2)\{s_1\}\{s_2\} \rightsquigarrow \text{split}(n_1,n_2)\{s_1\}\{s'_2\} + \eta',\sigma',\Sigma'} \text{ S-SPLIT-RIGHT} \\
\\
\frac{\eta,\sigma,\Sigma,t,b,p \text{ mod } n \vdash^{(h,n)} s \rightsquigarrow s' + \eta',\sigma',\Sigma'}{\eta,\sigma,\Sigma,t,b,p \vdash^{(h,q \cdot n)} \text{group } q \ s \rightsquigarrow \text{group } q \ s' + \eta',\sigma',\Sigma'} \text{ S-GROUP} \\
\\
\frac{\eta,\sigma,\Sigma,t,b,t \text{ mod } T \vdash^{(\texttt{Thread},T)} s \rightsquigarrow s' + \eta',\sigma',\Sigma'}{\eta,\sigma,\Sigma,t,b,0 \vdash^{(\texttt{Block},1)} \text{destruct in } s \rightsquigarrow \text{destruct in } s' + \eta',\sigma',\Sigma'} \text{ S-DESTRUCT-BLOCK} \\
\\
\frac{\eta,\sigma,\Sigma,t,b,p \vdash^{(\texttt{Block},1)} x := \text{alloc Shared } \tau \ n \ \text{in } s \rightsquigarrow s + \eta,\sigma[x \mapsto^\pi \langle x, n \rangle], \Sigma}{\eta,\sigma,\Sigma,t,b,p \vdash^{(\texttt{Block},1)} x := \text{alloc Shared } \tau \ n \ \text{in } s \rightsquigarrow s + \eta',\sigma',\Sigma} \text{ S-ALLOC-SHARED}
\end{array}$$

Fig. 15. Core semantic rules of Bundl. As with the typing rules, we present only a simplified selection of the full rules, which can be found in Appendix A.3.

This represents a minor difference from Prism, which uses additional static analysis to ensure that a claimed variable is only accessed in a single **split** branch.

Lastly, the **T-LOWER** rule mirrors the **T-DESTRUCT** rule; it uses the \downarrow operator to move a variable from one level of the hierarchy to another, distributing it equally among all the narrower perspectives at that level in the same manner as **T-PARTITION**.

Note that these rules only apply to thread-shared memory, i.e., arrays that do not live in **Local**. The provenance of an array type is denoted by the superscript l above it.

4.2 Bundl Semantics

Having explained the key rules of the type system, we can move on to discuss Bundl’s operational semantics. To reflect the fact that a GPU program executes in parallel across numerous threads, we model the semantics of Bundl in the style of Turon et al. [57], using a two-level small step judgment. We present the key rules of this semantics in Figure 15.

The top level (i.e., machine-level) judgment has just one rule: **S-PROGRAM**. This rule acts as a “frame” for the lower level (i.e., thread-level) judgment, and steps a collection of thread-ID-indexed local memories (L), a collection of block-ID-indexed shared memories (S), a global memory (Σ), and a thread pool (P) to an updated collection of memories and updated thread pool. The thread pool maps thread and block IDs to code, intuitively representing the program being executed by each thread at the current moment. The **S-PROGRAM** rule non-deterministically chooses a thread ID and block ID and steps it according to the thread-level judgment. This allows the semantics to model the full range of non-deterministic behavior arising from the GPU’s thread scheduler.⁶

⁶In reality, the GPU’s warp scheduler issues instructions to threads in a warp in lockstep, but modeling every thread as completely independent is both simpler and a conservative overestimate of the nondeterministic behavior of the GPU.

The thread-level judgment has the shape $\eta, \sigma, \Sigma, t, b, p \vdash^\pi s \rightsquigarrow s' \dashv \eta', \sigma', \Sigma'$, where

- η denotes thread-local memory,
- σ denotes shared memory,
- Σ denotes global memory,
- t denotes the thread’s ID,
- b denotes the ID of the block in which the thread lives, and
- p denotes the relative position of the thread within π (the perspective ID).

Critically, notice that a π also appears on the thread-level judgment just as it does on the typing judgment. This is because the thread semantics *dynamically tracks and enforces perspectives*. The same way evaluation of a program “gets stuck” if a value does not have the right type, the semantics of Bundl also get stuck if code attempts to access data or invoke commands with the wrong perspective. As an example, observe the **S-ALLOC-SHARED** rule in Figure 15, which requires a (**Block**, 1) perspective and will fail to step if encountered with a different one. This runtime perspective is present in Bundl, but is erased by Prism during compilation; in Section 4.3 we use it to prove that well-typed programs will always execute with the same perspective that the type system viewed them with.

The semantic rules for perspectives involve manipulating p to track which threads take which code paths when perspectives are **split** or **grouped**. Notice that in **S-PROGRAM**, the thread-level judgment always begins with perspective (**Grid**, 1): all the perspective management rules are congruences, narrowing the perspective of further evaluation as determined by the particular rule used.

These rules take great care to ensure that p always describes the relative position of a compute resource within its perspective; the payoff is that Bundl’s semantics can later use this p value to model the way that Prism automatically adjusts indices into data when **partitioning** a data perspective. Beyond these key rules for managing perspective on code, we have modeled all other core features of Prism, such as asynchronous operations and thread synchronization, in Bundl. To handle such features we equip the operational semantics with additional structure, including sets of semaphores [25] for synchronization and a stack of effect handlers for modeling deferred asynchronous computations inspired by Ahman and Pretnar [3]. We have elided these details here for simplicity, but interested readers can find them in their full complexity in Appendix A.3.

4.3 Soundness Theorem

Together, the type system and operational semantics allow us to prove the following syntactic soundness theorem, which says that Bundl programs are type safe and do not get stuck trying to execute operations for which they lack the required perspective:

THEOREM 4.1. *(Type-and-Perspective Safety). For any program s such that $\Gamma \vdash^\pi s$, either:*

- (1) s is **skip**, or
- (2) *for any well-typed environments η, σ , and Σ , there is an $s', \eta', \sigma', \text{ and } \Sigma'$ such that $\eta, \sigma, \Sigma, t, b, p \vdash^\pi s \rightsquigarrow^* s' \dashv \eta', \sigma', \Sigma'$ and $\Gamma' \vdash^\pi s'$, where Γ' is an extension of Γ , and $\eta', \sigma', \text{ and } \Sigma'$ are well-typed with respect to Γ' .*

PROOF. Via the usual progress and preservation lemmas, available in Appendix A.4. \square

It is worth noting that this soundness theorem guarantees a syntactic safety property, not a liveness property: it does not guarantee that all threads sharing perspective π that *can* reach a program point typed with π *will* eventually do so. Indeed, in the presence of nontermination, liveness does not hold—some of the threads could **split** off and loop forever. While we believe the liveness version of this theorem holds for a terminating fragment of Bundl, it is not provable with syntactic methods; the proof would require semantic techniques that are notoriously challenging and would be a research contribution [5, 28, 57] in and of itself. We plan to tackle this proof for Bundl in future work.

5 Implementation

Prism is implemented as an embedded language in Python. Once a program type checks, Prism lowers it to a CUDA file. All perspective information is erased during this step, and the generated CUDA contains no run-time checks. The file can then be compiled by **nvcc** [40], NVIDIA’s closed-source compiler, to produce an executable. Because Prism operates at roughly the same level of abstraction as CUDA, there is a one-to-one mapping between most language constructs and their CUDA counterparts. A notable change is the addition of three parameters to each device function: the thread’s relative ID, the block’s ID, and an offset for shared memory allocations.

Inserting Synchronization. Most of Prism’s implementation is straightforward, but inserting synchronization points is more involved. As described in Section 3.4, once data has been **partitioned**, Prism is responsible for synchronizing the data after the **partition** ends.

To determine where this synchronization must occur, Prism constructs a *data-control-flow graph* from the program. Nodes correspond to **partitions**, and edges capture program-order precedence: a parent **partition** must complete before its child begins. The graph can have backedges introduced through loops. In this graph, each **partition** is categorized as a read or a write **partition** by checking whether the **partitioned** variable ever appears as an lvalue. Synchronization points are inserted according to the following scheme:

- (1) If the *parent partition* is a write, a synchronization point is inserted before the current one to ensure that it observes the most recent data; or
- (2) If the *current partition* is a write, a synchronization point is inserted before it to ensure that all preceding reads have completed.

Figure 16 shows an example graph with synchronization points derived from these two conditions. The inferred synchronization points in Figure 6 have also been marked on lines 40, 48, 56, and 63.

Using this information, Prism emits *wait* operations before **partitions** begin and *arrive* operations after they end, using CUDA’s general *split-barrier* [20] primitive to implement them. For special cases, such as synchronizing an entire block or warp, Prism instead uses primitives like `__syncthreads()` or `__syncwarp()`.

Synchronization for asynchronous data movement is handled in exactly the same way and uses the same underlying graph. CUDA allows asynchronous loads to be associated with a split barrier, so Prism binds each asynchronous transfer to the appropriate wait–arrive pair inferred from the graph. Certain features, such as commit group-style synchronization [17, 18], require additional reasoning, and Prism performs further static analysis to insert the necessary synchronization.

It is worth noting that naively inserting synchronization immediately after each **partition** would be correct but prohibitively slow. To avoid this, Prism applies two optimizations: a wait-motion pass pushes waits downward toward the first use of the **partitioned** variable, and an arrive-motion pass pulls arrives upward toward its last use.

6 Evaluation

Having explained the design of Prism, we now evaluate it in the context of three main questions:

- RQ1** Can Prism express a variety of composable CUDA programs?
- RQ2** Can Prism express programs that use advanced GPU features?
- RQ3** Can Prism match the performance of existing, speed-of-light CUDA code?

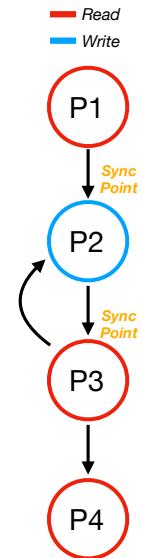


Fig. 16. An example data-control-flow graph with synchronization points.

To perform this evaluation, we use two GPUs. The first is the NVIDIA H100 SXM5, a server-grade chip that supports Tensor Core operations and a dedicated hardware copy engine, the Tensor Memory Accelerator (TMA) [16]. Notably, the H100 introduces a new logical level called the warpgroup, and we show that our programming model can accommodate it. Moreover, because the H100 has historically served as the primary GPU for large-scale AI training, many CUDA kernels are already highly optimized for this hardware and achieve near speed-of-light performance, providing a rigorous baseline for comparison. To ensure our results generalize beyond the H100, we also test programs on a second GPU, the NVIDIA 4070 SuperTi—a consumer-grade chip.

As mentioned in Section 5, when Prism typechecks a program, it produces a CUDA file. We compile the CUDA file with **nvcc** version 12.3 [40] with flags **-O3 -use_fast_math**. We initialize all inputs using a random number generator [31] and report the average runtime sampled over 10 iterations, following a warm-up phase of 5 iterations. All aggregate results reported in this section are geometric means. All programs in the evaluation have been reproduced in Appendix B.

6.1 RQ1: Can Prism Express a Variety of CUDA Programs?

We evaluate Prism’s expressivity by writing programs that have fundamentally different patterns of convergence, along with a library modeled after CUB [38] that contains composable pieces.

6.1.1 Programs with Different Convergence Behavior.

Matrix Multiplication (4070Ti). We chose matrix multiplication as our first benchmark for two main reasons. First, there exist several implementations that achieve near-peak performance, providing a strong baseline. Second, it allows for a range of increasingly sophisticated implementations that exercise different parts of the language, making it ideal for evaluating expressiveness.

We adapt the codebase from Boehm [6] to implement **float** matrix multiplication, commonly referred to as **sgemm**. Concretely, we compute $C \leftarrow \alpha AB + \beta C$ where A , B and C are matrices and α and β are scalars. As this is a **float** benchmark, it does not need advanced GPU features like asynchrony or Tensor Cores to achieve speed-of-light performance. We will discuss these in Section 6.2.

We implement five variants of the **sgemm** benchmark: (1) a naive version that follows the traditional single-program multiple-data pattern, written from the perspective of a thread; (2) a version that exploits memory coalescing [10], still expressed at the thread level; (3) a version that builds on (2) by introducing 2D tiling and staging data in shared memory, which requires shifting first to the perspective of a block and then to the perspective of a thread, while also requiring block-level synchronization; (4) a version that applies 2D tiling with vectorized loads, again written from the perspective of a block; and (5) a version that combines the optimizations from (2) through (4) while adding an additional level of tiling from the perspective of a warp.

We can express all variants cleanly, and the resulting programs are close to their CUDA counterparts, which distinguish perspectives through disciplined style. Prism, on the other hand, enforces this discipline at compile time. We evaluate the performance of these variants in Section 6.3.

Single-Pass Parallel Prefix Scan with Decoupled Look-Back (4070Ti). We also implement scan, a widely used parallel primitive, in Prism. We focus on the prefix-sum scan, which computes, for each position in an array, the sum of all elements up to that position. Prefix sum sits in a different corner of the GPU design space from matrix multiplication: it is memory intensive, requires careful attention to the convergence behavior of threads, and traditionally requires multiple passes over data.

We implement the single-pass parallel prefix scan with decoupled look-back, introduced by Merrill and Garland [36], an elegant algorithm that does not require multiple passes over the input data, and involves several distinct points of convergence. Within each block, work is decomposed into fine-grained thread-level and warp-level scans. After producing the local result, blocks publish their

partial prefix to global memory. Finally, each block waits until enough information from earlier blocks is available, at which point it accumulates the value and completes its section of the scan.

We implement this full strategy in Prism. Our implementation uses **unsafe** to implement a global-memory spinlock that lets a block check when the previous block's data is ready.

6.1.2 Case Study: Can Prism help build composable functions?

While answering the other RQs, we found ourselves developing a small library of functions—similar in spirit to CUB—that we would frequently call. In this section, we qualitatively study how Prism can help programmers design libraries that they can compose with confidence.

As mentioned in Section 1, CUB occupies a unique design space in the GPU library ecosystem. Unlike many other libraries such as cuBLAS [12], cuDNN [14], and cuSPARSE [15], which provide host-side functions, CUB provides a device-side library organized into different levels. It makes these levels apparent by prefixing each of its functions with **Device**, **Block**,

```

1 template<typename T, int BlockDimX,
2         int ItemsPerThread, BlockLoadAlgorithm Algorithm,
3         int BlockDimY = 1, int BlockDimZ = 1>
4 class BlockLoad:
5
6     // --- Calling the load function by instantiating the class ---
7
8     using BlockLoad = cub::BlockLoad<int, 128, 4, BLOCK_LOAD_DIRECT>;
9
10    // Allocate shared memory for BlockLoad
11    __shared__ typename BlockLoad::TempStorage temp_storage;
12    int thread_data[4]; // Thread local data
13
14    BlockLoad(temp_storage).Load(d_data, thread_data);

```

Fig. 17. Using **BlockLoad** in CUB.

and **Thread**. The prefix sum already used variants of these functions, translated into Prism.

Let's turn our attention to a particular CUB function—**BlockLoad**—and examine how it is equivalently expressed in Prism; we will see how Prism's type system reifies CUB's implicit assumptions. In CUB, the load function is implemented as a class, as shown in Figure 17.

CUB exposes a leaky abstraction, where information about the number of threads, block sizes, and other details seeps through:

- (1) The CUB documentation needs to specify the number of threads that the function can assume to be available, because within the function, each thread must locate itself in the computation and use its **threadIdx.x** accordingly.
- (2) The “item per thread” design can serve two purposes. The first is performance: if loops have constant bounds, they can be unrolled. The second is correctness: the function relies on the assumption that all threads call the function with an equal number of values to load.
- (3) The CUB documentation specifies that **thread_data** can be data local to each thread.
- (4) Finally, it says that if shared memory is being overwritten, a **__syncthreads()** call must be made to ensure that all reads have completed.

In Prism, however, we are able to describe these requirements through the interface shown in Figure 18, reducing the need to communicate numerous implementation details through documentation:

```

1 @prism("device")
2 @requires(block[1], thread[32])
3 def block_load(input : ptr(const int) @ block[1],
4               output : ptr(int) @ thread[1],
5               items_per_thread : int @ block[1]):

```

Fig. 18. **block_load** signature in Prism.

- (1) We do not need to pass in the number of threads at all. Whenever Prism calls a function, it tracks the relative thread ID, so each function can be written locally as if it were running alone, rather than having to determine where the thread resides on the grid.
- (2) We do not need to make **item_per_thread** a template argument for *correctness*. Its frequency is set at the function signature, so Prism will never allow a function to be called with a value living at a narrower perspective.
- (3) In our interface, **thread_data** is explicitly set to a thread-local value. Since it is not marked as **const**, Prism conservatively assumes it may be written to, and enforces at compile time that only **thread[1]** values are passed in.

- (4) Finally, using our synchronization pass outlined in Section 5, a `__syncthreads()` call will be inserted automatically if `input` is going to be used for writing.

Moreover, our version of CUB’s function is built as a composition of two other functions, whose interfaces are shown in Figures 19 and 20. The `block_load` function is implemented by a call to `warp_load`, which in turn calls `thread_load`. Had we mistakenly attempted to call `warp_load` from inside `thread_load`, however, Prism would reject this, rather than silently failing at runtime.

6.2 RQ2: Can Prism Express Programs that Use Advanced GPU Features?

To answer this question, we write a matrix multiplication for the `bf16` datatype on the H100, also known as `hgemm`. This benchmark is an acid test of our language, as `hgemm` pushes several language features to the extreme. To write an `hgemm` that can hit peak throughput on an H100, we need to write a warp-specialized kernel that uses the TMA—an asynchronous hardware copy engine that can move tiles of data at a time—and the warpgroup-level Tensor Core instructions, or `wmma` [9]—new to the Hopper architecture. A high-performance kernel for this matrix-multiplication overlaps computation with data movement by pipelining loads.

The implementation in Prism looks different from CUDA code, particularly in how pipelining is expressed. Since Prism uses named variables introduced by `partitions` or `claims` to determine the synchronization each region requires, when pipelining, we cannot dynamically change the pipeline slot simply by maintaining an index that wraps around based on the pipeline’s length. Instead, each pipeline slots must be given separate names so that Prism can track them independently and overlap compute with data-movement. This leads to pipeline slots that must be individually named and forces the load logic to be effectively “unrolled”. This, in turn, forces all pipelines in Prism to be statically sized. In practice, these pipelines are statically sized anyway to ensure they fit in shared memory.

Notably, for this benchmark, in addition to `wmma` and TMA, the program needs to dynamically reallocate registers between producer and consumer warpgroups, an instruction only available on Hopper. This redistribution is a warpgroup-level collective operation, and Prism can check it like any other. Moreover, getting `wmma` to work did not require introducing a new perspective into the language; `thread[128]` was sufficient. We did, however, need to add a dedicated TMA-style asynchronous data-movement construct, since Prism must eventually insert the appropriate synchronization for these transfers.

6.3 RQ3: Can Prism Match the Performance of Existing, Speed-of-Light CUDA Code?

In Section 6.1 and Section 6.2, we examined programs that expressed the same computation in multiple ways, relied on multiple points of convergence, and used advanced GPU features. We now discuss their performance.

The performance of the matrix multiplication benchmarks is shown in Figure 21, where we demonstrate that Prism is competitive with cuBLAS [12].

```

1 @prism("device")
2 @requires(thread[32])
3 # block_load calls into warp_load
4 def warp_load(input : ptr(const(int)) @ thread[32],
5               output : ptr(int) @ thread[1],
6               items_per_thread : int @ thread[32]):
```

Fig. 19. `warp_load` signature in Prism.

```

1 @prism("device")
2 @requires(thread[1])
3 # warp_load calls into thread_load
4 def thread_load(input : ptr(const(int)) @ thread[1],
5                  output : ptr(int) @ thread[1],
6                  items_per_thread : int @ thread[1]):
```

Fig. 20. `thread_load` signature in Prism.

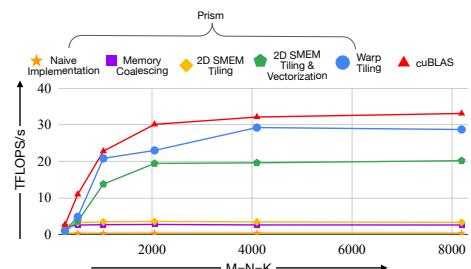


Fig. 21. Performance of `sgemm` on square matrices as matrix dimension $M=N=K$ increases.

For the prefix scan, we compare our performance to CUB in Figure 22, which shows that Prism is within 7% of CUB’s achieved bandwidth for arrays that do not fit in the L2 cache.

Finally, we evaluate our H100 implementation and show (in Figure 23) that it is competitive with cuBLAS on square sizes, coming within 15%. This performance difference arises due to one of `nvcc`’s optimization passes failing to trigger: the dynamic register allocation instruction is merely a hint—not a directive—to `nvcc`, and its optimization pass is sensitive to the way that pipeline structure is expressed. We emphasize that this is an exacting benchmark, and coming close to cuBLAS’s performance demonstrates Prism’s ability to control low-level features.

7 Related Work

Prism builds on a rich tradition of systems languages for GPU programming and theoretical foundations of parallel programming. Prism differs from these systems in a crucial way: it treats collective operations, and therefore composability, as first-class concerns.

Imperative Languages for GPUs. CUDA [39], ROCm [2], and OpenCL [53] are imperative programming languages that expose low-level access to GPU hardware. These languages offer no support for managing collective operations, statically or otherwise; instead, users must orchestrate computation on the machine by describing how individual threads execute code. The subtle failures permitted by this model are the motivation for Prism.

Descend [35], a new Rust-inspired [50] language, uses a type system to track aspects of the compute and memory hierarchy and is thus the closest to our work. However, Descend’s main concern is preventing data races, as opposed to ensuring that collective operations execute in valid contexts. As a result, Descend lacks support for many collective operations like Tensor Cores; in fact, because Descend allows threads to read their own IDs and thus induce data-dependent control flow, adding such support is not possible with its current design. It is also worth noting that while Descend does formalize a type system, it does not attempt to prove any properties about it.

Functional Languages for GPUs. Futhark [33], Accelerate [7] and Vertigo [27] are functional array languages with compilers targeting CUDA. These languages expose a high level interface, abstracting away details of the hardware entirely in exchange for stronger safety guarantees. Prism, however, exposes low-level details of the hardware and thus does not trade performance for safety.

Tile-Based Kernel DSLs. More recently, tile-based GPU languages—Triton [55], Pallas [52], Tilus [26], and Helion [44]—offer a middle ground between high-level abstraction and low-level control. However, these languages sidestep the question of composability entirely because they restrict programmers to a single layer of the hierarchy: a block (Triton, Tilus, Helion), or a warpgroup (Pallas). Gluon [56], meanwhile, is a new low-level, tile-based language, but it does not check and cannot enforce that collective operations are executed at the correct layer of the hierarchy.

Task-Based and Scheduling Languages. Languages like Cypress [58], Halide [45], Fireiron [29] and RISE/ELEVATE [30] expose a scheduling language that lets users modify an existing reference

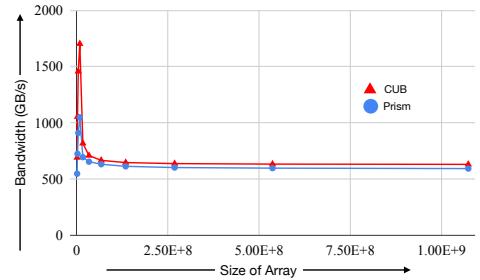


Fig. 22. Performance for prefix sum as input array size increases.

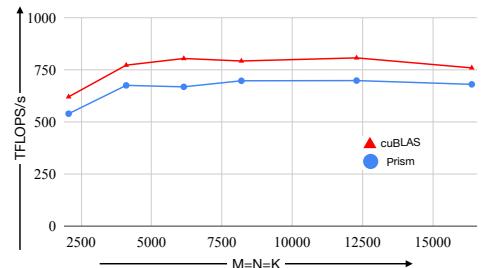


Fig. 23. Performance of `hgemm` on square matrices as matrix dimension $M=N=K$ increases.

program through external commands. Because these languages operate by transforming a fixed source of truth, they expose a fundamentally different programming model than Prism.

Libraries Built on CUDA. Many GPU libraries, such as CUTLASS [51], CUB [38], and ThunderKittens [47], expose device functions that operate at various levels of the compute hierarchy. These functions are typically organized via C++ namespaces to reflect their intended usage (e.g., warp-level, block-level). However, this organization is purely conventional: it encodes hierarchy through naming rather than enforcing it statically. As a result, correct use requires careful discipline from both the library implementer (to uphold naming and usage invariants) and the user (to correctly interpret them). Any mismatch or subtle misunderstanding between the intended and actual use of these functions goes unchecked by the compiler.

Theoretical Foundations. The design of Bundl is heavily inspired by existing work on dependency tracking [1]. Dependency tracking calculi allow type systems to track how data and code depend on each other, and have commonly been used to implement secure information flow analyses [24]. In Bundl, data that lives at a narrow perspective is unable to flow into data living at a broader perspective, and we use dependency tracking to capture this restriction in Bundl’s type system.

Prior theoretical work has tackled reasoning about thread divergence. In particular, Muller and Hoffmann [37] build a sophisticated quantitative program logic for this use case. Singhania [46], meanwhile, uses static analysis techniques to predict thread divergence to unlock optimizations.

8 Conclusion, Limitations, and Future Work

We have presented Prism, a new, low-level GPU language that statically guarantees safe usage of compute resources by construction, without sacrificing low-level control. Prism introduces a new mental model for writing GPU code, which we are excited to make more expressive and ergonomic.

Particularly, we plan to improve the experience of writing pipelines in Prism. As discussed in Section 6.2, Prism currently requires pipeline slots to be explicitly named; we can make this more ergonomic this by adding language support for generating pipeline-style code.

Prism can be extended to more architectures; in particular, it can accommodate newer GPUs—such as Blackwell—that support coarser-grained Tensor Core operations. More broadly, we hope to generalize Prism’s model to other hierarchical compute environments, including distributed systems.

We also believe that Prism is capable of guaranteeing data-race freedom, but additional work is required to support this both formally and in practice. In particular, we think that if Prism restricted users to **partitions** with injective indexing functions, data-race freedom would follow naturally. We are also interested in exploring how the design principles of Descend [35], which build on Rust’s ownership model [50], could be applied to Prism.

On the theoretical side, we plan to explore a terminating fragment of Bundl and prove the liveness property discussed in Section 4.3: that all threads sharing perspective π eventually reach the parts of a program viewed at that π . This amounts to showing that threads sharing the same perspective execute the same code and observe the same data, which we hope to prove using logical relations following Turon et al. [57] and Spies et al. [48, 49].

We believe that Prism is a promising low-level substrate that enables confident composition and can serve as a foundation for building higher-level libraries and abstractions.

9 Acknowledgements

We would like to thank Ege Kabasakaloglu for writing the single-pass parallel prefix scan with decoupled look-back in Prism, Amanda Liu for early conversations about Prism’s formalization, and Adam Chlipala for helping us crystalize the **partition** construct. Paul Biberstein, Christophe Gyurgyik, Olivia Hsu, Fredrik Kjolstad, Scott Kovach, Benjamin Pierce, Alexander Root, Adrian

Sampson, Shiv Sundram, Stephanie Weirich, Rohan Yadav, and Bobby Yan provided helpful feedback on many drafts of the paper.

References

- [1] Martín Abadi, Anindya Banerjee, Nevin Heintze, and Jon G. Riecke. [n. d.]. A core calculus of dependency. In *Proceedings of the 26th ACM SIGPLAN-SIGACT symposium on Principles of programming languages* (San Antonio Texas USA, 1999-01). ACM, 147–160. <https://doi.org/10.1145/292540.292555>
- [2] Advanced Micro Devices, Inc. 2024. AMD ROCm™ Software. <https://www.amd.com/en/products/software/rocm.html>. Accessed: 2025-11-10.
- [3] Danel Ahman and Matija Pretnar. 2021. Asynchronous effects. *Proceedings of the ACM on Programming Languages* 5, POPL (Jan. 2021), 1–28. <https://doi.org/10.1145/3434305>
- [4] Michael Bauer, Sean Treichler, and Alex Aiken. 2014. Singe: leveraging warp specialization for high performance on GPUs. *SIGPLAN Not.* 49, 8 (Feb. 2014), 119–130. <https://doi.org/10.1145/2692916.2555258>
- [5] Lars Birkedal, Filip Sieczkowski, and Jacob Thamsborg. 2012. A Concurrent Logical Relation. In *Computer Science Logic (CSL'12) - 26th International Workshop/21st Annual Conference of the EACSL (2012)*. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 107–121. <https://doi.org/10.4230/LIPIcs.CSL.2012.107>
- [6] Simon Boehm. 2022. SGEMM_CUDA: Fast CUDA matrix multiplication from scratch. https://github.com/siboehm/SGEMM_CUDA. GitHub repository. Accessed: 2025-11-10.
- [7] Manuel M.T. Chakravarty, Gabriele Keller, Sean Lee, Trevor L. McDonell, and Vinod Grover. 2011. Accelerating Haskell array codes with multicore GPUs. In *Proceedings of the Sixth Workshop on Declarative Aspects of Multicore Programming* (Austin, Texas, USA) (DAMP '11). Association for Computing Machinery, New York, NY, USA, 3–14. <https://doi.org/10.1145/1926354.1926358>
- [8] NVIDIA Corporation. 2022. NVIDIA H100 GPU Whitepaper. (2022). <https://resources.nvidia.com/en-us-hopper-architecture/nvidia-h100-tensor-c> Accessed: 2025-11-12.
- [9] NVIDIA Corporation. 2025. Asynchronous Warpgroup Level Matrix Multiply-Accumulate Instructions. <https://docs.nvidia.com/cuda/parallel-thread-execution/#asynchronous-warpgroup-level-matrix-instructions> Accessed: 2025-11-13.
- [10] NVIDIA Corporation. 2025. Coalesced Access to Global Memory. <https://docs.nvidia.com/cuda/cuda-c-best-practices-guide/#coalesced-access-to-global-memory> Accessed: 2025-11-13.
- [11] NVIDIA Corporation. 2025. cub::BlockReduce. https://nvidia.github.io/ccl/cub/api/classcub_1_1BlockReduce.html Accessed: 2025-11-09.
- [12] NVIDIA Corporation. 2025. cuBLAS. <https://docs.nvidia.com/cuda/cublas/index.html> Accessed: 2025-11-09.
- [13] NVIDIA Corporation. 2025. cuBLASDx. <https://docs.nvidia.com/cuda/cublasdx/#nvidia-cublasdx> Accessed: 2025-11-11.
- [14] NVIDIA Corporation. 2025. cuDNN. <https://docs.nvidia.com/deeplearning/cudnn/latest/> Accessed: 2025-11-09.
- [15] NVIDIA Corporation. 2025. cuSPARSE. <https://docs.nvidia.com/cuda/cusparse/index.html> Accessed: 2025-11-09.
- [16] NVIDIA Corporation. 2025. Data Movement and Conversion Instructions: Bulk Copy. <https://docs.nvidia.com/cuda/parallel-thread-execution/#data-movement-and-conversion-instructions-bulk-copy> Accessed: 2025-11-12.
- [17] NVIDIA Corporation. 2025. Data Movement and Conversion Instructions: cp.async.bulk.commit_group. <https://docs.nvidia.com/cuda/parallel-thread-execution/#data-movement-and-conversion-instructions-cp-async-bulk-commit-group> Accessed: 2025-11-13.
- [18] NVIDIA Corporation. 2025. Data Movement and Conversion Instructions: cp.async.commit_group. <https://docs.nvidia.com/cuda/parallel-thread-execution/#data-movement-and-conversion-instructions-cp-async-commit-group> Accessed: 2025-11-13.
- [19] NVIDIA Corporation. 2025. Data Movement and Conversion Instructions: Non-bulk Copy. <https://docs.nvidia.com/cuda/parallel-thread-execution/#data-movement-and-conversion-instructions-non-bulk-copy> Accessed: 2025-11-12.
- [20] NVIDIA Corporation. 2025. Parallel Synchronization and Communication Instructions: mbarrier. <https://docs.nvidia.com/cuda/parallel-thread-execution/#parallel-synchronization-and-communication-instructions-mbarrier> Accessed: 2025-11-13.
- [21] NVIDIA Corporation. 2025. Synchronization Functions. <https://docs.nvidia.com/cuda/cuda-c-programming-guide/#synchronization-functions> Accessed: 2025-11-11.
- [22] NVIDIA Corporation. 2025. Warp Shuffle Functions. <https://docs.nvidia.com/cuda/cuda-c-programming-guide/#warp-shuffle-functions> Accessed: 2025-11-10.
- [23] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni,

Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhua Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanja Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhihang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. DeepSeek-V3 Technical Report. <https://doi.org/10.48550/arXiv.2412.19437> arXiv:2412.19437.

- [24] Dorothy E. Denning and Peter J. Denning. 1977. Certification of programs for secure information flow. *Commun. ACM* 20, 7 (July 1977), 504–513. <https://doi.org/10.1145/359636.359712>
- [25] Edsger Wybe Dijkstra. 1963. Over de sequentialiteit van procesbeschrijvingen. (1963). <https://training-ir7.tdl.org/handle/123456789/594>
- [26] Yaoya Ding, Bohan Hou, Xiao Zhang, Allan Lin, Tianqi Chen, Cody Yu Hao, Yida Wang, and Gennady Pekhimenko. 2025. Tulus: A Tile-Level GPGPU Programming Language for Low-Precision Computation. arXiv:2504.12984 [cs.LG] <https://arxiv.org/abs/2504.12984>
- [27] C. Elliott. 2005. Vertigo – GPU Compiler & Embedded Language for Graphics Processors. <http://conal.net/Vertigo/>. Accessed: 2025-11-10.
- [28] Azadeh Farzan, Zachary Kincaid, and Andreas Podelski. 2016. Proving Liveness of Parameterized Programs. In *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science*. ACM, New York NY USA, 185–196. <https://doi.org/10.1145/2933575.2935310>
- [29] Bastian Hagedorn, Archibald Samuel Elliott, Henrik Barthels, Rastislav Bodik, and Vinod Grover. 2020. Fireiron: A Data-Movement-Aware Scheduling Language for GPUs. In *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques* (Virtual Event, GA, USA) (PACT ’20). Association for Computing Machinery, New York, NY, USA, 71–82. <https://doi.org/10.1145/3410463.3414632>
- [30] Bastian Hagedorn, Johannes Lenfers, Thomas Kundefinedhler, Xueying Qin, Sergei Gorlatch, and Michel Steuwer. 2020. Achieving high-performance the functional way: a functional pearl on expressing high-performance optimizations as rewrite strategies. *Proc. ACM Program. Lang.* 4, ICFP, Article 92 (Aug. 2020), 29 pages. <https://doi.org/10.1145/3408974>
- [31] Horace He. 2024. Strangely, Matrix Multiplications on GPUs Run Faster When Given “Predictable” Data! <https://www.thonking.ai/p/strangely-matrix-multiplications> Accessed: 2025-11-13.
- [32] Troels Henriksen, Niels G. W. Serup, Martin Elsman, Fritz Henglein, and Cosmin E. Oancea. 2017. Futhark: purely functional GPU-programming with nested parallelism and in-place array updates. In *Proceedings of the 38th ACM SIGPLAN Conference on Programming Language Design and Implementation*. ACM, Barcelona Spain, 556–571. <https://doi.org/10.1145/3062341.3062354>
- [33] Troels Henriksen, Niels G. W. Serup, Martin Elsman, Fritz Henglein, and Cosmin E. Oancea. 2017. Futhark: purely functional GPU-programming with nested parallelism and in-place array updates. *SIGPLAN Not.* 52, 6 (June 2017), 556–571. <https://doi.org/10.1145/3140587.3062354>
- [34] Jan Hoffmann. 2011. *Types with potential: polynomial resource bounds via automatic amortized analysis*. Ph. D. Dissertation. lmu.
- [35] Bastian Köpcke, Sergei Gorlatch, and Michel Steuwer. [n. d.]. Descend: A Safe GPU Systems Programming Language. 8 ([n. d.]), 841–864. Issue PLDI. <https://doi.org/10.1145/3656411>
- [36] Duane Merrill and Michael Garland. 2016. *Single-pass Parallel Prefix Scan with Decoupled Look-back*. Technical Report NVR-2016-002. NVIDIA Corporation. Accessed: 2025-11-10.
- [37] Stefan K. Muller and Jan Hoffmann. [n. d.]. Modeling and analyzing evaluation cost of CUDA kernels. 5 ([n. d.]), 1–31. Issue POPL. <https://doi.org/10.1145/3434306>
- [38] NVIDIA Corporation. 2025. *CUB: CUDA Unbound*. <https://docs.nvidia.com/cuda/cub/index.html> CUDA Toolkit Documentation.

- [39] NVIDIA Corporation. 2025. *CUDA C++ Programming Guide*. NVIDIA. <https://docs.nvidia.com/cuda/cuda-c-programming-guide/> Accessed: 2025-11-06.
- [40] NVIDIA Corporation. 2025. CUDA Toolkit 12.3 Downloads (Archive). <https://developer.nvidia.com/cuda-12-3-0-download-archive>. Accessed: 2025-11-11.
- [41] NVIDIA Corporation. 2025. Parallel Thread Execution (PTX) ISA – Asynchronous Warpgroup Level Matrix Multiply-Accumulate Instructions. <https://docs.nvidia.com/cuda/parallel-thread-execution/index.html?highlight=tngen05%2520cp#asynchronous-warpgroup-level-matrix-instructions>. Accessed: 2025-11-10.
- [42] NVIDIA Corporation. 2025. Parallel Thread Execution (PTX) ISA – TensorCore 5th Generation Instructions (tcgen05). <https://docs.nvidia.com/cuda/parallel-thread-execution/index.html?highlight=tngen05%2520cp#tensorcore-5th-generation-instructions>. Accessed: 2025-11-10.
- [43] NVIDIA Corporation. 2025. Parallel Thread Execution (PTX) ISA – Warp Level Matrix Instructions. <https://docs.nvidia.com/cuda/parallel-thread-execution/index.html?highlight=tngen05%2520cp#warp-level-matrix-instructions>. Accessed: 2025-11-10.
- [44] PyTorch Team at Meta. 2025. Helion: A High-Level DSL for Performant and Portable ML Kernels. <https://pytorch.org/blog/helion/>. Accessed: 2025-11-10.
- [45] Jonathan Ragan-Kelley, Connelly Barnes, Andrew Adams, Sylvain Paris, Frédéric Durand, and Saman Amarasinghe. 2013. Halide: a language and compiler for optimizing parallelism, locality, and recomputation in image processing pipelines. *SIGPLAN Not.* 48, 6 (June 2013), 519–530. <https://doi.org/10.1145/2499370.2462176>
- [46] Nimit Singhania. 2018. *Static Analysis for GPU Program Performance*. PhD dissertation. University of Pennsylvania, Philadelphia, PA, USA. Supervisors: Rajeev Alur and Joseph Devietti.
- [47] Benjamin Frederick Spector, Simran Arora, Aaryan Singhal, Arjun Parthasarathy, Daniel Y Fu, and Christopher Re. 2025. ThunderKittens: Simple, Fast, and \$textit{Adorable} Kernels. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=0fJfVOSUra>
- [48] Simon Spies, Lennard Gähler, Daniel Gratzner, Joseph Tassarotti, Robbert Krebbers, Derek Dreyer, and Lars Birkedal. 2021. Transfinite Iris: resolving an existential dilemma of step-indexed separation logic. In *Proceedings of the 42nd ACM SIGPLAN International Conference on Programming Language Design and Implementation*. ACM, Virtual Canada, 80–95. <https://doi.org/10.1145/3453483.3454031>
- [49] Simon Spies, Neel Krishnaswami, and Derek Dreyer. 2021. Transfinite step-indexing for termination. *Proceedings of the ACM on Programming Languages* 5, POPL (Jan. 2021), 1–29. <https://doi.org/10.1145/3434294>
- [50] Rust Team. 2025. Rust Programming Language. <https://rust-lang.org/>
- [51] Vijay Thakkar, Pradeep Ramani, Cris Cecka, Aniket Shivam, Honghao Lu, Ethan Yan, Jack Kosaian, Mark Hoemmen, Haicheng Wu, Andrew Kerr, Matt Nicely, Duane Merrill, Dustyn Blasig, Aditya Atluri, Fengqi Qiao, Piotr Majcher, Paul Springer, Markus Hohnerbach, Jin Wang, and Manish Gupta. 2023. CUTLASS. <https://github.com/NVIDIA/cutlass>
- [52] The JAX Authors. 2024. Pallas:Mosaic GPU – A JAX Kernel Language for GPUs. <https://docs.jax.dev/en/latest/pallas/gpu/index.html>. Accessed: 2025-11-10.
- [53] The Khronos Group Inc. 2025. OpenCL™ – The Open Standard for Parallel Programming of Heterogeneous Systems. <https://www.khronos.org/opencl/>. Accessed: 2025-11-10.
- [54] Philippe Tillet. 2025. Introducing Triton: Open-source GPU programming for neural networks. <https://openai.com/index/triton/>
- [55] Philippe Tillet, H. T. Kung, and David Cox. 2019. Triton: an intermediate language and compiler for tiled neural network computations. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on Machine Learning and Programming Languages* (Phoenix, AZ, USA) (MAPL 2019). Association for Computing Machinery, New York, NY, USA, 10–19. <https://doi.org/10.1145/3315508.3329973>
- [56] triton-lang. 2025. “01-intro.py” – Introduction to Gluon tutorial in Triton. <https://github.com/triton-lang/triton/blob/main/python/tutorials/gluon/01-intro.py>. Accessed: 2025-11-10.
- [57] Aaron J. Turon, Jacob Thamsborg, Amal Ahmed, Lars Birkedal, and Derek Dreyer. [n. d.]. Logical relations for fine-grained concurrency. In *Proceedings of the 40th annual ACM SIGPLAN-SIGACT symposium on Principles of programming languages* (Rome Italy, 2013-01-23). ACM, 343–356. <https://doi.org/10.1145/2429069.2429111>
- [58] Rohan Yadav, Michael Garland, Alex Aiken, and Michael Bauer. 2025. Task-Based Tensor Computations on Modern GPUs. *Proc. ACM Program. Lang.* 9, PLDI, Article 163 (June 2025), 25 pages. <https://doi.org/10.1145/3729262>

A Complete Bundl Type System, Semantics, and Syntactic Soundness Proofs

A.1 Basic Definitions

$$\begin{aligned}
 & \text{Hierarchy Levels } h ::= \mathbf{Grid} \mid \mathbf{Block} \mid \mathbf{Thread} \\
 & \text{Memory Kinds } l ::= \mathbf{Local} \mid \mathbf{Shared} \mid \mathbf{Global} \\
 & \text{Perspectives } \pi : h \times \mathbb{N} \\
 & \text{Base Types } b ::= \mathbf{bool} \mid \mathbf{int} \mid \mathbf{float} \\
 & \text{Types } \tau ::= b \mid b[]^l \mid \mathbf{Fun}(\Gamma, \pi, m) \mid \mathbf{async} \tau \\
 & \text{Contexts } \Gamma ::= \cdot \mid \Gamma, x : \pi \tau \\
 & \text{Shared Memory Remaining } m : \mathbb{N}
 \end{aligned}$$

Perspectives π are part of an algebra parameterized over some constant values T (the number of threads per block) and B (the number of blocks per grid). With these values, we have two isomorphisms:

$$(\mathbf{Block}, 1) \cong (\mathbf{Thread}, T)$$

and

$$(\mathbf{Grid}, 1) \cong (\mathbf{Block}, B)$$

The **group** and **split** operations of Prism allows us to move along this isomorphism, from left to right. For clarity, in Bundl we split these operations into three: a **split** operation that can split perspectives into multiple narrower ones, and a **destruct** operation that directly moves us along the isomorphism, and a **group** that divides our current perspective into equal sized parts.

Perspectives π are also lexicographically ordered in the obvious way. hs are ordered such that $\mathbf{Thread} \leq \mathbf{Block} \leq \mathbf{Grid}$, and $(h_1, n_1) \leq (h_2, n_2)$ iff $n_1 | n_2$ and $h_1 \leq h_2$.

We define scalar multiplication $i \times h$ of natural numbers with hs : $i \times (h, n) = (h, in)$.

We also define division of perspectives and hierarchy levels of type $\pi \times \pi \rightarrow \mathbb{N}$. $\mathbf{Grid}/\mathbf{Block} = B$ and $\mathbf{Block}/\mathbf{Thread} = T$. We lift this to perspectives like so: $(h_1, n_1) / (h_2, n_2) = ((h_1/h_2) \cdot n_1) / n_2$.

Lastly we define a partial \downarrow operator on hs such that $\downarrow \mathbf{Block} = \mathbf{Thread}$ and $\downarrow \mathbf{Grid} = \mathbf{Block}$. Note that $\downarrow \mathbf{Thread}$ is undefined. This operator lifts to π s whose second component is 1 and encodes the leftward component of the isomorphism above: $\downarrow (\mathbf{Block}, 1) = (\mathbf{Thread}, T)$ and $\downarrow (\mathbf{Grid}, 1) = (\mathbf{Block}, B)$.

Note also that the presentation of these rules in the main body of the paper elide the m portion, which tracks the maximum amount of memory a given computation is allowed to use. In the full system presented here, both the typing rules and the operational semantics carry an additional piece of information tracking allocated memory.

A.2 Complete Bundl Typing Rules

A.2.1 Expressions.

$$\frac{x : \pi \tau \in \Gamma}{\Gamma \vdash^\pi x : \tau} \text{-Var} \quad \frac{}{\Gamma \vdash^\pi n : \mathbf{int}} \text{-Int} \quad \frac{}{\Gamma \vdash^\pi f : \mathbf{float}} \text{-Float}$$

$$\frac{}{\Gamma \vdash^\pi b : \mathbf{bool}} \text{-Bool} \quad \frac{\pi < (\mathbf{Grid}, 1)}{\Gamma \vdash^\pi \mathbf{partition_id} : \mathbf{int}} \text{-Partition-Id}$$

$$\frac{\Gamma \vdash^\pi' e_1 : \tau[]^l \quad \Gamma \vdash^\pi e_2 : \mathbf{int} \quad l = \mathbf{Global} \text{ or } l = \mathbf{Local} \quad \pi \leq \pi'}{\Gamma \vdash^\pi e_1[e_2] : \tau} \text{-Arr-Access}$$

$$\begin{array}{c}
 \frac{\Gamma \vdash^{\pi'} e_1 : \tau[] \text{ Shared} \quad \Gamma \vdash^{\pi} e_2 : \mathbf{int} \quad \pi \leq (\mathbf{Block}, 1) \quad \pi \leq \pi'}{\Gamma \vdash^{\pi} e_1[e_2] : \tau} \text{ T-Arr-Access-Shared} \\
 \\
 \frac{\Gamma \vdash^{\pi} e_1 : \mathbf{int} \quad \Gamma \vdash^{\pi} e_2 : \mathbf{int}}{\Gamma \vdash^{\pi} e_1 \mathbf{bop} e_2 : \mathbf{int}} \text{ T-Bop} \quad \frac{\Gamma \vdash^{\pi} e_1 : \mathbf{int} \quad \Gamma \vdash^{\pi} e_2 : \mathbf{int}}{\Gamma \vdash^{\pi} e_1 \mathbf{cmp} e_2 : \mathbf{bool}} \text{ T-Cmp}
 \end{array}$$

A.2.2 Statements.

$$\begin{array}{c}
 \frac{f : \pi \mathbf{Fun}(x_i : \tau_i, \pi, m') \in \Gamma \quad \Gamma \vdash^{\pi} e_i : \tau_i \quad m' \leq m}{\Gamma \vdash_m^{\pi} f(e_1, \dots, e_n)} \text{ T-Function-Call} \\
 \\
 \frac{\Gamma \vdash_m^{(h, n_1)} s_1 \quad \Gamma \vdash_m^{(h, n_2)} s_2 \quad n_1, n_2 \text{ align to } n}{\Gamma \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{s_1\}\{s_2\}} \text{ T-Split} \quad \frac{\Gamma \vdash_m^{\downarrow \pi} s}{\Gamma \vdash_m^{\pi} \mathbf{destruct} \text{ in } s} \text{ T-Destruct} \\
 \\
 \frac{\Gamma \vdash_m^{(h, n)} s}{\Gamma \vdash_m^{(h, q \cdot n)} \mathbf{group} q s} \text{ T-Group} \quad \frac{x : \pi' \tau \in \Gamma \quad \Gamma \vdash^{\pi} e : \tau \quad \pi' \leq \pi}{\Gamma \vdash_m^{\pi} x = e} \text{ T-Assn} \\
 \\
 \frac{}{\Gamma \vdash_m^{\pi} \mathbf{init}_{\psi}} \text{ T-Sync-Init} \quad \frac{}{\Gamma \vdash_m^{\pi} \mathbf{dec}_{\psi}} \text{ T-Sync-Dec} \\
 \\
 \frac{}{\Gamma \vdash_m^{\pi} \mathbf{wait}_{\psi}} \text{ T-Sync-Wait} \quad \frac{}{\Gamma \vdash_m^{\pi} \mathbf{skip}} \text{ T-Skip} \quad \frac{n \leq m}{\Gamma \vdash_m^{\pi} \mathbf{free} n} \text{ T-Free} \\
 \\
 \frac{\Gamma \vdash^{\pi} e : \tau \quad \Gamma, x : \pi' \tau \vdash_m^{\pi} s \quad \pi' \leq \pi \quad \tau \text{ not an array type}}{\Gamma \vdash_m^{\pi} x : \tau @ \pi' = e \text{ in } s} \text{ T-Decl} \\
 \\
 \frac{\Gamma \vdash^{\pi'} e_1 : \tau[]^l \quad \Gamma \vdash^{\pi} e_2 : \mathbf{int} \quad \Gamma \vdash^{\pi} e_3 : \tau \quad l = \mathbf{Global} \text{ or } l = \mathbf{Local} \quad \pi' \leq \pi}{\Gamma \vdash_m^{\pi} e_1[e_2] = e_3} \text{ T-Arr-Assn} \\
 \\
 \frac{\Gamma \vdash^{\pi'} e_1 : \tau[] \text{ Shared} \quad \Gamma \vdash^{\pi} e_2 : \mathbf{int} \quad \Gamma \vdash^{\pi} e_3 : \tau \quad \pi \leq (\mathbf{Block}, 1) \quad \pi' \leq \pi}{\Gamma \vdash_m^{\pi} e_1[e_2] = e_3} \text{ T-Arr-Assn-Shared} \\
 \\
 \frac{\Gamma \vdash^{\pi} e : \mathbf{bool} \quad \Gamma \vdash_m^{\pi} s_1 \quad \Gamma \vdash_m^{\pi} s_2}{\Gamma \vdash_{\max(m_1, m_2)}^{\pi} \mathbf{if} e \text{ then } s_1 \text{ else } s_2} \text{ T-If} \\
 \\
 \frac{\Gamma \vdash^{\pi} e : \mathbf{bool} \quad \Gamma \vdash_m^{\pi} s}{\Gamma \vdash_m^{\pi} \mathbf{while} e \text{ do } s} \text{ T-While} \quad \frac{\Gamma \vdash_m^{\pi} s_1 \quad \Gamma \vdash_m^{\pi} s_2}{\Gamma \vdash_{\max(m_1, m_2)}^{\pi} s_1; s_2} \text{ T-Seq} \\
 \\
 \frac{\Gamma, x : \pi \tau[]^l \vdash_m^{\pi} s \quad l = \mathbf{Global} \text{ or } l = \mathbf{Local}}{\Gamma \vdash_{m+n \cdot \text{size}(\tau)}^{\pi} x := \mathbf{alloc} \ l \ \tau \ n \text{ in } s} \text{ T-Alloc} \\
 \\
 \frac{\Gamma, x : (\mathbf{Block}, 1) \tau[] \text{ Shared} \vdash_m^{\pi} s}{\Gamma \vdash_{m+n \cdot \text{size}(\tau)}^{\pi} x := \mathbf{alloc} \ Shared \ \tau \ n \text{ in } s} \text{ T-Alloc-Shared} \\
 \\
 \frac{\Gamma, y : (h, n/c) \tau[]^l \vdash_m^{(h, n)} s \quad c | n \quad l \neq \mathbf{Local}}{\Gamma, x : (h, n) \tau[]^l \vdash_m^{(h, n)} \mathbf{partition}_{\psi} x \text{ into } y \text{ by } c \text{ in } s} \text{ T-Partition} \\
 \\
 \frac{\Gamma, y : (h, n') \tau[]^l \vdash_m^{(h, n')} s \quad n' \leq n \quad l \neq \mathbf{Local}}{\Gamma, x : (h, n) \tau[]^l \vdash_m^{(h, n)} \mathbf{claim}_{\psi} x \text{ into } y \text{ at } n' \text{ in } s} \text{ T-Claim}
 \end{array}$$

$$\begin{array}{c}
\frac{\Gamma, y : \downarrow^\pi \tau[]^l \vdash_m^\pi s \quad l \neq \text{Local}}{\Gamma, x : {}^\pi \tau[]^l \vdash_m^\pi \text{lower}_\psi x \text{ into } y \text{ in } s} \text{-Lower} \\
\\
\frac{\Gamma, y : {}^{(\text{Thread},1)} \text{async } \tau[]^l \vdash_m^{(\text{Thread},1)} s}{\Gamma, x : {}^{(\text{Thread},1)} \tau[]^l \vdash_m^{(\text{Thread},1)} \text{async_partition}_\phi x \text{ into } y \text{ in } s} \text{-Async-Partition} \\
\\
\frac{}{\Gamma, x : {}^{(\text{Thread},1)} \text{async } \tau[], y : {}^{(\text{Thread},1)} \tau[]^l' \vdash_m^{(\text{Thread},1)} \text{async_memcpy}(x,y)} \text{-Async-Memcpy} \\
\\
\frac{}{\Gamma, x : {}^\pi \tau[]^l, y : {}^\pi \tau[]^l' \vdash_m^\pi \text{memcpy}(x,y)} \text{-Memcpy}
\end{array}$$

A.3 Complete Bundl Semantics

A.3.1 Definitions.

Global Memory $\Sigma ::= \cdot | \Sigma, n \mapsto^\pi v$

Shared Memory $\sigma ::= \cdot | \sigma, n \mapsto^\pi v$

Local Memory $\eta ::= \cdot | \eta, n \mapsto^\pi v$

Block Memory Map $S ::= \forall n \in B, n \mapsto \sigma$

Thread Memory Map $L ::= \forall n \in T, n \mapsto \eta$

Synchronization Map $\Psi : \psi \rightarrow \mathbb{N} \rightarrow \mathbb{N}$

Deferred Computations Map $\Phi : \phi \rightarrow \{s\}$

In real GPUs, thread IDs are only unique within their block. However, in this calculus for simplicity we assume thread IDs are global. One can convert back and forth between this abstracted notion of a thread ID and a block-unique ID via addition modulo T . That is, $t_{\text{real}} = t_{\text{simplified}} \bmod T$ and $t_{\text{simplified}} = t_{\text{real}} + b \cdot T$.

By convention the names for local and shared and global memory do not conflict, as on the GPU they will be separate pointer spaces. Additionally, we freely interchange between using names for variables and integer locations.

In the main body of the paper, for simplicity we elide the synchronization map and the deferred computation map from the operational semantics, as our theorems do not make any guarantees about non-interference. However, as they are part of the full semantics, we include them here for completeness. By convention the synchronization and deferred computation maps are a total functions, initialized to map to $\lambda_. 0$ for ψ s and $\lambda_. \{\}$ for ϕ s not explicitly initialize.

The shape of the judgment for a single thread is $\eta, \sigma, \Sigma, t, b, p, \Psi \vdash_m^\pi s \rightsquigarrow s' \dashv_m \eta', \sigma', \Sigma', \Psi'$. The t here is the thread ID, the b is the block ID, and the p is the perspective ID. The last of these three is modified and managed by the rules for **split**, **group** and **destruct**, and tracks the relative position of the thread within a group. This semantics is in a small step style.

The shape of the judgment for expressions is $\eta, \sigma, \Sigma \vdash_\pi^\pi e \Downarrow v$. The two π s represent the ambient compute context (i.e., the context in which resources are being read), while π' , represents the target compute context (i.e., the compute context of the variable into which the result of the expression is going to be written). This is relevant for computing the value of **partition_id**, which divides the two contexts. As a shorthand, we can divide a perspective by a scalar value like so: $(h,n)/c = (h,n)/(h,c)$.

The overall evaluation of a program is expressed as

$$L, S, \Sigma, P, \Psi, \Phi \rightsquigarrow L', S', \Sigma', P', \Psi', \Phi'.$$

In this judgment P serves as a *thread pool*, mapping pairs of thread and block IDs (which don't change) to statements and memory (which can be updated by stepping). One can think of P as tracking which program is running on each thread. This steps according to the following rule:

$$\frac{L(t), S(b), \Sigma, t, b, 0, \Psi, \Phi \vdash_m^{(\text{Grid}, 1)} s \rightsquigarrow s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi' \quad P(t, b) = (s, m)}{L, S, \Sigma, P, \Psi, \Phi \rightsquigarrow L[t \mapsto \eta'], S[b \mapsto \sigma'], \Sigma', P[(t, b) \mapsto (s', m')], \Psi', \Phi'} \text{ S-Program}$$

For simplicity of notation, we define an **update** operation that searches the three environments for the one that contains the variable being used (by convention, there is no conflict between the environments, as in reality they exist in three separate address spaces). We also define a similar **get** operation that retrieves a variable from memory, and a **rename** operation that remaps a variable with the same value but under a different name.

$$\begin{aligned} \mathbf{update}(\eta, \sigma, \Sigma, x, v) &= (\eta[x \mapsto^\pi v], \sigma, \Sigma) \text{ when } x \in^\pi \eta \\ \mathbf{update}(\eta, \sigma, \Sigma, x, v) &= (\eta, \sigma[x \mapsto^\pi v], \Sigma) \text{ when } x \in^\pi \sigma \\ \mathbf{update}(\eta, \sigma, \Sigma, x, v) &= (\eta, \sigma, \Sigma[x \mapsto^\pi v]) \text{ when } x \in^\pi \Sigma \\ \\ \mathbf{get}(\eta, \sigma, \Sigma, x) &= \eta(x) \text{ when } x \in^\pi \eta \\ \mathbf{get}(\eta, \sigma, \Sigma, x) &= \sigma(x) \text{ when } x \in^\pi \sigma \\ \mathbf{get}(\eta, \sigma, \Sigma, x) &= \Sigma(x) \text{ when } x \in^\pi \Sigma \\ \\ \mathbf{rename}(\eta, \sigma, \Sigma, x, y, \pi') &= (\eta[y \mapsto^{\pi'} \eta(x)], \sigma, \Sigma) \text{ when } x \in^\pi \eta \\ \mathbf{rename}(\eta, \sigma, \Sigma, x, y, \pi') &= (\eta, \sigma[y \mapsto^{\pi'} \sigma(x)], \Sigma) \text{ when } x \in^\pi \sigma \\ \mathbf{rename}(\eta, \sigma, \Sigma, x, y, \pi') &= (\eta, \sigma, \Sigma[y \mapsto^{\pi'} \Sigma(x)]) \text{ when } x \in^\pi \Sigma \end{aligned}$$

A.3.2 Perspective Management Rules.

$$\frac{p < n_1 \quad n_1, n_2 \text{ align to } n \quad \eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n_1)} s_1 \rightsquigarrow s'_1 \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{s_1\}\{s_2\} \rightsquigarrow \mathbf{split}(n_1, n_2)\{s'_1\}\{s_2\} \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{ S-Split-Left}$$

$$\frac{p < n_1 \quad n_1, n_2 \text{ align to } n}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{\mathbf{skip}\}\{s_2\} \rightsquigarrow \mathbf{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Split-Left-Done}$$

$$\frac{p \geq n_1 \quad p < n_1 + n_2 \quad n_1, n_2 \text{ align to } n \quad \eta, \sigma, \Sigma, t, b, p - n_1, \Psi, \Phi \vdash_m^{(h, n_2)} s_2 \rightsquigarrow s'_2 \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{s_1\}\{s_2\} \rightsquigarrow \mathbf{split}(n_1, n_2)\{s_1\}\{s'_2\} \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{ S-Split-Right}$$

$$\frac{p \geq n_1 \quad p < n_1 + n_2 \quad n_1, n_2 \text{ align to } n}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{s_1\}\{\mathbf{skip}\} \rightsquigarrow \mathbf{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Split-Right-Done}$$

$$\frac{p \geq n_1 + n_2 \quad n_1, n_2 \text{ align to } n}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n)} \mathbf{split}(n_1, n_2)\{s_1\}\{s_2\} \rightsquigarrow \mathbf{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Split-None}$$

$$\begin{array}{c}
\frac{\eta, \sigma, \Sigma, t, b, t \text{ mod } T, \Psi, \Phi \vdash_m^{(\text{Thread}, T)} s \rightsquigarrow s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, 0, \Psi, \Phi \vdash_m^{(\text{Block}, 1)} \text{destruct in } s \rightsquigarrow \text{destruct in } s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{S-Destruct-Block} \\
\\
\frac{\eta, \sigma, \Sigma, t, b, b \text{ mod } B, \Psi, \Phi \vdash_m^{(\text{Block}, B)} s \rightsquigarrow s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, 0, \Psi, \Phi \vdash_m^{(\text{Grid}, 1)} \text{destruct in } s \rightsquigarrow \text{destruct in } s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{S-Destruct-Grid} \\
\\
\frac{}{\eta, \sigma, \Sigma, t, b, 0, \Psi, \Phi \vdash_m^{\pi} \text{destruct in skip} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{S-Destruct-Done} \\
\\
\frac{\eta, \sigma, \Sigma, t, b, p \text{ mod } n, \Psi, \Phi \vdash_m^{(h, n)} s \rightsquigarrow s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, q \cdot n)} \text{group } q \ s \rightsquigarrow \text{group } q \ s'; \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{S-Group} \\
\\
\frac{}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{group } q \ \text{skip} \rightsquigarrow \text{skip}; \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{S-Group-Done}
\end{array}$$

A.3.3 Thread Synchronization. We define a **size** operation on perspectives to compute the size of a perspective (the number of individual compute resources sharing it). The operation is defined thusly:

$$\begin{aligned}
\text{size(Thread, } n) &= n \\
\text{size(Block, } n) &= n \cdot T \\
\text{size(Grid, } n) &= n \cdot B \cdot T
\end{aligned}$$

$$\begin{array}{c}
\frac{\Psi(\psi)(p) = 0}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{wait}_{\psi} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{S-Sync-Wait-Done} \\
\\
\frac{\Psi(\psi)(p) \neq 0}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{wait}_{\psi} \rightsquigarrow \text{wait}_{\psi} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{S-Sync-Wait-Spin} \\
\\
\frac{\Psi' = \Psi(\psi)[p \mapsto \Psi(\psi)(p) - 1]}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{dec}_{\psi} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi', \Phi} \text{S-Sync-Dec} \\
\\
\frac{\Psi(\psi)(p) = 0 \quad \Psi' = \Psi(\psi)[p \mapsto \text{size}(\pi)]}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{init}_{\psi} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi', \Phi} \text{S-Sync-Init-Zero} \\
\\
\frac{\Psi(\psi)(p) \neq 0}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{init}_{\psi} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{S-Sync-Init-Nonzero}
\end{array}$$

A.3.4 Asynchrony.

$$\begin{array}{c}
\frac{\text{rename}(\eta, \sigma, \Sigma, x, y, (\text{Thread}, 1)), t, b, p, \Psi, \Phi \vdash_m^{(\text{Thread}, 1)} s \rightsquigarrow s' \dashv_{m'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{m'}^{(\text{Thread}, 1)} \text{async_partition}_{\phi} \ x \ \text{into } y \ \text{in } s \rightsquigarrow \text{async_partition}_{\phi} \ x \ \text{into } y \ \text{in } s' \dashv_m \eta', \sigma', \Sigma', \Psi', \Phi'} \text{S-Async-Partition-Congr} \\
\\
\frac{\Phi = \Phi'[\phi \mapsto \Phi'(\phi) \cup \{s\}]}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(\text{Thread}, 1)} \text{async_partition}_{\phi} \ x \ \text{into } y \ \text{in skip} \rightsquigarrow (\text{async_partition}_{\phi} \ x \ \text{into } y \ \text{in } s) \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi'} \text{S-Async-Partition-Unwind}
\end{array}$$

$$\begin{array}{c}
\frac{\Phi(\phi) = \emptyset}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\text{(Thread,1)}} \text{async_partition}_\phi x \text{ into } y \text{ in skip} \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Async-Partition-Done} \\
\\
\frac{}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\text{(Thread,1)}} \text{async_memcpy}(x, y) \rightsquigarrow \text{skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi[\phi \mapsto \Phi(\phi) \cup \{\text{memcpy}(x, y)\}]} \text{ S-Async-Memcpy} \\
\\
\frac{(\eta', \sigma', \Sigma') = \text{update}(\eta, \sigma, \Sigma, x, \text{get}(\eta, \sigma, \Sigma, y))}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi \text{memcpy}(x, y) \rightsquigarrow \text{skip} \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Memcpy} \\
\\
\text{A.3.5} \quad \text{Variables and Memory.} \\
\\
\frac{\eta, \sigma, \Sigma \vdash_{\pi'}^\pi e \Downarrow v \quad \pi' \leq \pi}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi x : \tau @ \pi' := e \text{ in } s \rightsquigarrow s \dashv_m \eta[x \mapsto^{\pi'} v], \sigma, \Sigma, \Psi, \Phi} \text{ S-Decl} \\
\\
\frac{}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi \text{free } n \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Free} \\
\\
\frac{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi x := \text{alloc Local } \tau n \text{ in } s \rightsquigarrow s; \text{ free } (n \cdot \text{size}(\tau)) \dashv_{m+n \cdot \text{size}(\tau)} \eta[x \mapsto^{\pi} \langle x, n \rangle], \sigma, \Sigma, \Psi, \Phi}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi x := \text{alloc Shared } \tau n \text{ in } s \rightsquigarrow s; \text{ free } (n \cdot \text{size}(\tau)) \dashv_{m+n \cdot \text{size}(\tau)} \eta, \sigma[x \mapsto^{\pi} \langle x, n \rangle], \sigma, \Sigma, \Psi, \Phi} \text{ S-Alloc-Local} \\
\\
\frac{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi x := \text{alloc Global } \tau n \text{ in } s \rightsquigarrow s; \text{ free } (n \cdot \text{size}(\tau)) \dashv_{m+n \cdot \text{size}(\tau)} \eta, \sigma, \Sigma[x \mapsto^{\pi} \langle x, n \rangle], \Psi, \Phi}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{\pi'}^\pi x = e \rightsquigarrow \text{skip} \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Alloc-Global} \\
\\
\frac{x \in^{\pi'} \eta, \sigma, \Sigma \quad \eta, \sigma, \Sigma \vdash_{\pi'}^\pi e \Downarrow v \quad (\eta', \sigma', \Sigma') = \text{update}(\eta, \sigma, \Sigma, v) \quad \pi' \leq \pi}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi, \Phi \vdash_m^\pi x = e \rightsquigarrow \text{skip} \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Assn} \\
\\
\frac{\eta, \sigma, \Sigma, t, b, p \vdash_\pi^\pi e_1 \Downarrow \langle l, i \rangle \quad i < n \quad \pi' \leq \pi \quad \eta, \sigma, \Sigma, t, b, p \vdash_\pi^\pi e_2 \Downarrow i \quad x \in^{\pi'} \eta, \sigma, \Sigma \quad \eta, \sigma, \Sigma \vdash_{\pi'}^\pi e_3 \Downarrow v \quad (\eta', \sigma', \Sigma') = \text{update}(\eta, \sigma, \Sigma, l+i, v)}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi e_1[e_2] = e_3 \rightsquigarrow \text{skip} \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Arr-Assn} \\
\\
\frac{s' = s[(y + c \cdot p) / y] \quad (\eta', \sigma', \Sigma') = \text{rename}(\eta, \sigma, \Sigma, x, y, \pi/c)}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi \text{partition}_\psi xycs \rightsquigarrow \text{init}_\psi; s'; \text{dec}_\psi; \text{wait}_\psi \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Partition} \\
\\
\frac{(\eta', \sigma', \Sigma') = \text{rename}(\eta, \sigma, \Sigma, x, (h, n_1))}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{(h, n_1 + n_2)} \text{claim}_\psi x \text{ into } y \text{ at } n_1 \text{ in } s \rightsquigarrow \text{init}_\psi; \text{split}(n_1, n_2); \{s'\}; \text{skip}; \text{dec}_\psi; \text{wait}_\psi \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Claim} \\
\\
\frac{(\eta', \sigma', \Sigma') = \text{rename}(\eta, \sigma, \Sigma, x, y, \downarrow \pi)}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi \text{lower}_\psi x \text{ into } y \text{ in } s \rightsquigarrow \text{init}_\psi; s; \text{dec}_\psi; \text{wait}_\psi \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi} \text{ S-Lower}
\end{array}$$

A.3.6 Control Flow.

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi}^{\pi} e \Downarrow \text{true}}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{if } e \text{ then } s_1 \text{ else } s_2 \rightsquigarrow s_1 \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-If-True}$$

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi}^{\pi} e \Downarrow \text{false}}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{if } e \text{ then } s_1 \text{ else } s_2 \rightsquigarrow s_2 \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-If-False}$$

$$\frac{}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{while } e \text{ do } s \rightsquigarrow \text{if } e \text{ then } (s; \text{while } e \text{ do } s) \text{ else skip} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-While}$$

$$\frac{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} s_1 \rightsquigarrow s'_1 \dashv_{m'}^{\pi'} \eta', \sigma', \Sigma', \Psi', \Phi'}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} s_1; s_2 \rightsquigarrow s'_1; s_2 \dashv_{m'}^{\pi'} \eta', \sigma', \Sigma', \Psi', \Phi'} \text{ S-Seq-First}$$

$$\frac{}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} \text{skip}; s_2 \rightsquigarrow s_2 \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi} \text{ S-Seq-Done}$$

$$\frac{\Sigma(f) = \{[x_1 : \tau_1, \dots, x_n : \tau_n], s\} \quad \sigma, \Sigma \vdash_{\pi}^{\pi} e_i \Downarrow v_i \quad m' \leq m}{\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^{\pi} f(e_1, \dots, e_n) \rightsquigarrow s \dashv_m \eta[x_i \mapsto v_i], \sigma, \Sigma, \Psi, \Phi} \text{ S-Function-Call}$$

A.3.7 Expressions.

$$\frac{\pi < (\text{Grid}, 1) \quad \pi' \leq \pi}{\eta, \sigma, \Sigma \vdash_{\pi}^{\pi}, \text{partition_id} \Downarrow \pi / \pi' - 1} \text{ E-Partition-Id}$$

$$\frac{}{\eta, \sigma, \Sigma \vdash_{\pi}^{\pi}, x \Downarrow \text{get}(\eta, \sigma, \Sigma, x)} \text{ E-Var}$$

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1 \Downarrow \langle l, n \rangle \quad \eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_2 \Downarrow i \quad i < n \quad \pi' \leq \pi}{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1[e_2] \Downarrow \text{get}(\eta, \sigma, \Sigma, l+i)} \text{ E-Arr-Access}$$

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, \vdash n \Downarrow n}{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, \vdash n \Downarrow n} \text{ E-Int} \quad \frac{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, \vdash b \Downarrow b}{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, \vdash b \Downarrow b} \text{ E-Bool}$$

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1 \Downarrow v_1 \quad \eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_2 \Downarrow v_2 \quad v = v_1 \text{ bop } v_2}{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1 \text{ bop } e_2 \Downarrow v} \text{ E-Bop}$$

$$\frac{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1 \Downarrow v_1 \quad \eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_2 \Downarrow v_2 \quad v = v_1 \text{ cmp } v_2}{\eta, \sigma, \Sigma \vdash_{\pi'}^{\pi}, e_1 \text{ cmp } e_2 \Downarrow v} \text{ E-Cmp}$$

A.4 Theorems and Proofs

Note that in this section we assume no out of bounds array accesses. In general Prism (and by extension Bundl) makes no guarantees about array out of bounds.

A.4.1 More definitions. As a premise to our type safety theorems, we need to assume we have a well-typed environment, written $\Gamma \vdash \eta, \sigma, \Sigma$. We define what this means inductively

$$\begin{array}{c} \frac{}{\eta, \sigma, \Sigma \vdash n : \text{int}} \text{ V-Int} \quad \frac{}{\eta, \sigma, \Sigma \vdash b : \text{bool}} \text{ V-Bool} \quad \frac{}{\eta, \sigma, \Sigma \vdash f : \text{float}} \text{ V-Float} \\ \\ \frac{\forall i < n, \eta, \sigma, \Sigma \vdash \text{get}(\eta, \sigma, \Sigma, x + i) : \tau}{\eta, \sigma, \Sigma \vdash \langle x, n \rangle : \tau[]^l} \text{ V-Array} \quad \frac{\Gamma, x_i : \pi \tau_i \vdash_m^{\pi} s \quad \Gamma \vdash \cdot, \cdot, \text{fns } \Sigma}{\eta, \sigma, \Sigma \vdash \{x_i : \tau_i, s\} : \text{Fun}(x_i : \tau_i, \pi, m)} \text{ V-Function} \\ \\ \frac{}{\vdash \eta, \sigma, \Sigma} \text{ G-Empty} \quad \frac{\eta(x) =^{\pi} v \quad \eta, \sigma, \Sigma \vdash v : \text{int}}{\Gamma, x : \pi \text{ int} \vdash \eta, \sigma, \Sigma} \text{ G-Int} \end{array}$$

$$\begin{array}{c}
\frac{\eta(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \mathbf{bool}}{\Gamma, x :^\pi \mathbf{bool} \vdash \eta, \sigma, \Sigma} \text{ G-Bool} \quad \frac{\eta(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \mathbf{float}}{\Gamma, x :^\pi \mathbf{float} \vdash \eta, \sigma, \Sigma} \text{ G-Float} \\
\\
\frac{\eta(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \tau[]}{\Gamma, x :^\pi \tau[] \text{Local} \vdash \eta, \sigma, \Sigma} \text{ G-Local} \quad \frac{\sigma(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \tau[]}{\Gamma, x :^\pi \tau[] \text{Shared} \vdash \eta, \sigma, \Sigma} \text{ G-Shared} \\
\\
\frac{\Sigma(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \tau[]}{\Gamma, x :^\pi \tau[] \text{Global} \vdash \eta, \sigma, \Sigma} \text{ G-Global} \quad \frac{\Sigma(x) =^\pi v \quad \eta, \sigma, \Sigma \vdash v : \mathbf{Fun}(\Gamma', \pi, m)}{\Gamma, x :^\pi \mathbf{Fun}(\Gamma', \pi, m) \vdash \eta, \sigma, \Sigma} \text{ G-Function}
\end{array}$$

We can prove a couple simple lemmas about well-typed environments under operations like **rename**, **update**, and **get**.

LEMMA A.1. (*Well-typed get*) If $\Gamma \vdash \eta, \sigma, \Sigma$ and $x :^\pi \tau \in \Gamma$ then $\eta, \sigma, \Sigma \vdash \mathbf{get}(\eta, \sigma, \Sigma, x) : \tau$.

LEMMA A.2. (*Well-typed rename*) If $\Gamma, x :^\pi \tau \vdash \eta, \sigma, \Sigma$ then $\Gamma, x :^\pi \tau, y :^\pi \tau' \vdash \mathbf{rename}(\eta, \sigma, \Sigma, x, y, \pi')$.

LEMMA A.3. (*Well-typed update*) If $\Gamma \vdash \eta, \sigma, \Sigma$ and $\eta, \sigma, \Sigma \vdash v : \tau$ then $\Gamma, x :^\pi \tau \vdash \mathbf{update}(\eta, \sigma, \Sigma, x, v)$.

We also define a well-formedness precondition on p with respect to π :

$$(h, n) \vdash p ::= p < n$$

We also define well-formedness for the async stack:

$$\Gamma \vdash \Phi ::= \forall \phi, s \in \Phi(\phi), \Gamma \vdash_m^{(\mathbf{Thread}, 1)} s$$

A.5 Proofs

LEMMA A.4. (*Expression Safety*) If $\Gamma \vdash^\pi e : \tau$ and $\Gamma \vdash \eta, \sigma, \Sigma$ and $\pi \vdash p$ and $\pi' \leq \pi$, then there is some v such that $\eta, \sigma, \Sigma \vdash_{\pi'} e \Downarrow v$ and $v : \tau$.

PROOF. This proof proceeds by induction on the typing relation for expressions. Despite the fact that this property implies termination, we do not need a logical relation to prove it because the expression language is very simple.

The **T-Int**, **T-Float**, and **T-Bool** cases are trivial, using the rules **E-Int**, **E-Bool** and **E-Float** to compute values. In the case for **T-Partition-Id**, the π premises of the typing rules and our assumption that $\pi' \leq \pi$ match the premises of the evaluation rule, so this rule is simple as well.

The cases for **T-Bop** and **T-Cmp** follow directly from the inductive hypotheses, assuming a valid and correctly implemented set of binary operators and comparators.

The only interesting cases are **T-Arr-Access** and **T-Arr-Access-Shared**.

In both cases our inductive hypotheses and inversion give us that $\pi' \leq \pi$, and e_1 evaluates to a $\langle x, n \rangle$, and that all the values between x and $x+n$ in the appropriate environment are typed at τ . We also know that e_2 evaluates to an integer i . We assume that all array accesses are in bounds, so $i < l$, which is sufficient to use the **E-Arr-Access** rule to complete this case, and the proof. \square

LEMMA A.5. (*Expression Determinism*) If $\eta, \sigma, \Sigma, t, b, p \vdash_{\pi'}^\pi e \Downarrow v_1$ and $\eta, \sigma, \Sigma, t, b, p \vdash_{\pi'}^\pi e \Downarrow v_2$ then $v_1 = v_2$.

PROOF. Straightforward by induction on the semantic derivation. \square

LEMMA A.6. (*Expression Well-Typedness*) If $\Gamma \vdash_{\pi'}^\pi e : \tau$ and $\Gamma \vdash \eta, \sigma, \Sigma$ and $\pi \vdash p$, and $\eta, \sigma, \Sigma, t, b, p \vdash_{\pi'}^\pi e \Downarrow v$, then $v : \tau$.

PROOF. By our expression safety lemma our well-typed expression must evaluate to a well-typed value v' . By our determinism lemma v' must be the same as v , so v is well-typed. \square

LEMMA A.7. (*Statement Progress*) If $\Gamma \vdash \eta, \sigma, \Sigma$ and $\Gamma \vdash_m^\pi s$ and $\pi \vdash p$, then either s is **skip** or there is some s' such that $\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi s \rightsquigarrow s' \vdash_{m''} \eta', \sigma', \Sigma, \Psi', \Phi'$

PROOF. This proceeds by induction on the typing derivation.

Perspective Management Rules.

- **Case T-Split:**

In this case we have by our assumption that $\pi \vdash p$ that $p < n$. We also have that n_1, n_2 **align to** n , so $n_1 + n_2 \leq n$. There are three cases to consider, then: when $p < n_1$, when $p \geq n_1$ and $p < n_1 + n_2$, and when $p \geq n_1 + n_2$.

In the first case, we have by our inductive hypothesis that s is either **skip** or that it can step in an (h, n_1) context. In the former case we can use the **S-Split-Left-Done** rule and in the latter we can use the **S-Split-Left rule**.

The second case is almost symmetric. The only additional work we have to do is to argue that $(h, n_2) \vdash p - n_1$, or equivalently that $p - n_1 < n_2$. This, however, is immediate from our assumption that $p < n_1 + n_2$.

In the last case, we just use the **S-Split-None** rule to step to **skip**.

- **Case T-Destruct**

By our inductive hypothesis, we know that s can step at $\downarrow \pi$ if $\downarrow \pi \vdash p$. The \downarrow operation is only defined at **(Block, 1)** or **(Grid, 1)**, so we only need to consider the cases where π is one of those.

In the former case p becomes $t \mathbf{mod} T$ while $\downarrow \pi$ is **(Thread, T)**. $t \mathbf{mod} T < T$ for any t so this satisfies the requirement that $\pi \vdash p$, which lets us use our inductive hypothesis: s is either **skip** or can step. If it can step, we can use this to satisfy the premise of **S-Destruct-Block** to step in this case. If it is **skip**, then we use the rule **S-Destruct-Done** to step instead.

The latter case is the same, except using the fact that $b \mathbf{mod} B < B$ and the **S-Destruct-Grid** rule.

- **Case T-Group**

In this case we have by assumption that $(h, q \cdot n) \vdash p$, i.e., that $p < q \cdot n$.

In this case we have our IH that if $(h, n) \vdash p'$ for some p' , then s is either **skip** or steps with (h, n) perspective with p' as our perspective ID.

We choose p' to be $p \mathbf{mod} n$. This is always $< n$, so $(h, n) \vdash p'$. This lets us use our IH to get that s is either **skip** (in which case we can use the **S-Group-Done** rule to step) or itself steps, which lets us use the **S-Group** rule to step.

Thread Synchronization Rules.

- **Case T-Sync-Wait**

$\Psi(\psi)(p)$ is either zero or it is not. In the former case we use the **S-Sync-Wait-Done** rule and in the latter we use **S-Sync-Wait-Spin**.

- **Case T-Sync-Dec**

We use the **S-Sync-Dec** rule to step.

- **Case T-Sync-Init**

We use the **S-Sync-Init-Zero** or **S-Sync-Init-Nonzero** rules depending on whether $\Psi(\psi)(p)$ is zero or not.

Asynchrony Rules.

- **Case T-Async-Partition**

In this case, we have via our IH that if $\Gamma, y : (\text{Thread}, 1) \text{ async } \tau[]^l \vdash \eta', \sigma', \Sigma'$, then we can either step s with **(Thread, 1)** perspective under environments η' , σ' , and Σ' , or s is **skip**.

We have by assumption that $\Gamma, x:(\text{Thread}, 1) \text{ async } \tau[]^l \vdash \eta, \sigma, \Sigma$. By our environment renaming lemma, this gives us what we need to use our IH with $(\eta', \sigma', \Sigma')$ as **rename** $(\eta, \sigma, \Sigma, x, y, (\text{Thread}, 1))$. Thus s either steps or is **skip**. In the former case we can step with **S-Async-Partition-Congr**, and in the latter we can use either **S-Async-Partition-Unwind** or **S-Async-Partition-Done** depending on whether $\Phi(\phi)$ is empty or not.

- Case **T-Async-Memcpy**

Immediate via use of the **S-Async-Memcpy** rule.

- Case **T-Memcpy**

Immediate via use of the **S-Memcpy** rule.

Memory Rules.

- Case **T-Decl**

Via our lemma about expression type safety and our hypothesis that e is well-typed, we obtain the premises necessary to use the **S-Decl** rule to step.

- Case **T-Arr-Assn**

Each of e_1 , e_2 and e_3 must evaluate to a well-typed value by the expression type safety lemma. In particular, both e_1 evaluates to some $\langle l, n \rangle$ and e_2 evaluates to some i . We assume all array accesses are in bounds, so this is sufficient to use the **S-Arr-Assn** rule to step.

- Case **T-Arr-Assn-Shared**

Same as previous case.

- Case **T-Free**

Trivial via the **S-Free** rule.

- Case **T-Partition**

Trivial via the **S-Partition** rule.

- Case **T-Claim**

Trivial via the **S-Claim** rule.

- Case **T-Lower**

Trivial via the **S-Lower** rule.

- Case **T-Alloc**

We assume that l is not **Shared**, so we can use the **S-Alloc-Local** or **S-Alloc-Global** rule, depending on whether l is **Local** or **Global**.

- Case **T-Alloc-Shared**

Trivial via the **S-Alloc-Shared** rule.

Control Rules.

- Case **T-Skip**

Trivial

- Case **T-While**

Trivial, all while loops step via the **S-While** rule

- Case **T-If**

By our proof of expression type safety, the expression e steps to either the boolean value true or false. We can thus use either the **S-If-True** or **S-If-False** rules to step.

- Case **T-Seq**

In this case we know by our IH that s_1 is either **skip** or can step. In the former case we use the **S-Seq-Done** rule and in the latter we use the **S-Seq-First** rule.

- Case **T-Function-Call**

In this case we know by our expression safety lemma that each of the arguments will evaluate to a well-typed value. We also have by assumption that f has a function type, which by inversion

on the **V-Function** rule tells us that it is a closure type. Additionally our assumption that $\Gamma \vdash \eta, \sigma, \Sigma$ tell us that Σ contains f at the same type that Γ does. These premises are sufficient to use the **S-Function-Call** rule. \square

LEMMA A.8. (*Statement Preservation*) *If $\Gamma \vdash \eta, \sigma, \Sigma$ and $\Gamma \vdash_m^\pi s$ and $\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{m'}^\pi s \rightsquigarrow s' \dashv_{m''} \eta', \sigma', \Sigma, \Psi', \Phi'$ and $\pi \vdash p$ and $m \geq m'$ and $\Gamma \vdash \Phi$, then there is some Γ' such that $\Gamma \subseteq \Gamma'$ and $\Gamma' \vdash_m^\pi s'$ and $\Gamma' \vdash \eta', \sigma', \Sigma'$ and $m \geq m''$ and $\Gamma' \vdash \Phi'$.*

PROOF. We proceed by induction on the derivation of $\Gamma \vdash_m^\pi s$.

Perspective Management Rules.

- **Case T-Split**

In this case we have by assumption that $(h, n) \vdash p$, $\Gamma \vdash \eta, \sigma, \Sigma$, and $\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{m'}^\pi \text{split}(n_1, n_2)\{s_1\}\{s_2\} \rightsquigarrow s' \dashv_{m''} \eta', \sigma', \Sigma, \Psi', \Phi'$. Our inductive hypotheses give us that if for any p , if $(h, n_1) \vdash p$ and s_1 steps with perspective ID p , or if $(h, n_2) \vdash p$ and s_2 steps with perspective ID p then their results are well typed.

By inversion on our semantic derivation, we are in one of 5 cases.

In the **S-Split-Left** case $p < n_1$ and s_1 steps to s'_1 . This is sufficient to tell us that s'_1 is well-typed and the output environments of that relation η', σ' , and Σ' are all well-typed by $\Gamma' \supseteq \Gamma$, and that the memory is properly bounded by the typing rules.

We can thus use the **T-Split-Left** rule to conclude that the result of this case is well-typed. The **S-Split-Left-Done** rule is trivial via the **T-Skip** rule.

The **Right** cases are symmetric, with the observation that when $p \geq n_1$ and $p < n_1 + n_2$ then $p - n_1 < n_2$.

The last **S-Split-None** rule is trivial via the **T-Skip** rule.

- **Case T-Destruct**

In this case we have by assumption that $\downarrow \pi$ is defined, so π is either **(Block,1)** or **(Grid,1)**. We also assume that $\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{m'}^\pi \text{destruct in } s \rightsquigarrow s' \dashv_{m''} \eta', \sigma', \Sigma, \Psi', \Phi'$. We also have by our inductive hypothesis that for any p such that $\downarrow \pi \vdash p$ and s'' such that s steps to s'' at p , then that step preserved well-typedness.

By inversion on the step relation, we are in one of three cases.

If the rule used **S-Destruct-Block**, then we know that s steps to s'' and p is $t \bmod T$. $(\text{Thread}, T) \vdash t \bmod T$ for any t , so we can use our inductive hypothesis to conclude that the s'' stepped to by s is well-typed, as are its environments and memory usage. The **T-Destruct** rule then gives us our desired goal.

The **S-Destruct-Grid** case proceeds similarly, while the **S-Destruct-Done** case is trivial.

- **Case T-Group**

In this case we have by assumption that

$\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_{m'}^{(h, q \cdot n)} \text{group } q \rightsquigarrow s' \dashv_{m''} \eta', \sigma', \Sigma, \Psi', \Phi'$. We have by our inductive hypothesis that s steps to s'' at some perspective ID p' and $(h, n) \vdash p'$, then s'' is well typed, as are the other outputs of that step.

By inversion on the step relation, we are in one of two cases. The **S-Group-Done** case is trivial, so we shall focus on the **S-Group** case. In this case we have that s steps to s'' at perspective ID $p \bmod n$. It is always the case that $(h, n) \vdash p \bmod n$ for any p , so we can use our inductive hypothesis to conclude that s'' is well-typed, as are its output environments and memory usage. From there, it is a simple application of the **T-Group** rule to conclude that **group** q s' is well-typed, and to finish the case.

Thread Synchronization Rules. These rules are all trivial: with one exception all thread synchronization primitives step to skip without changing environment or memory, and are thus obviously well-typed. **S-Sync-Wait-Spin** does not produce skip, but it steps to the same statement as we already assumed typechecks in the premise of the lemma, so is straightforward nonetheless.

If we wanted to say something about deadlock freedom we'd have more work here, but we aren't doing that, so these rules are easy.

Asynchrony Rules.

- Case **T-Async-Partition** In this case we have that s is well-typed in a context where x has been renamed into y , with $(\text{Thread}, 1)$ perspective. We also have that $\Gamma, x : (\text{Thread}, 1) \vdash \tau[]^l \vdash \eta, \sigma, \Sigma$ and $\Gamma, x : (\text{Thread}, 1) \vdash \tau[]^l \vdash \Phi$.

By inversion, we are in one of three cases.

In the **S-Async-Partition-Done** case, we are done.

In the **S-Async-Partition-Congr** case, our inductive hypothesis gives us that there is some Γ' such that $\Gamma, y : (\text{Thread}, 1) \text{ async } \tau[]^l \subseteq \Gamma'$ and $\Gamma' \vdash \Phi'$ and $\Gamma' \vdash \text{rename}(\eta, \sigma, \Sigma, x, y, (\text{Thread}, 1))$ via our well-typed renaming lemma. This lets us use the **T-Async-Partition** rule to check this case, with a choice of Γ' as $\Gamma, x : (\text{Thread}, 1) \vdash \tau[]^l$.

In the **S-Async-Partition-Unwind** case, our assumption that Φ is well-typed tells us that $\Gamma, y : (\text{Thread}, 1) \vdash \{\ (\text{Thread}, 1)_m s$. Thus, we can use the **T-Async-Partition** rule to type this case.

- Case **T-Async-Memcpy**

In this case the statement and environment typing are trivial, we need only to show that the `async` stack remains well typed.

In this case we have that x and y have the same type at $(\text{Thread}, 1)$. This is sufficient for us to check $x = y$ at $(\text{Thread}, 1)$, meaning that adding that instruction to the stack maintains its well-typedness.

- Case **T-Memcpy**

Immediate via use of the **S-Memcpy** rule. We just need to show that the environment remains well typed, which we know via our lemmas about **update** and **get**.

Memory Rules.

- Case **T-Decl**

By inversion, we are using **S-Decl** rule for evaluation. Our well-typed expression lemma gives us that v is well-typed, so it follows from our assumptions and our lemmas about extending environments that the extended η and s are well-typed by $\Gamma, x : \pi \vdash \tau$.

- Case **T-Free**

Trivial.

- Case **T-Alloc**

By inversion we are either in the **S-Alloc-Local** or **S-Alloc-Global** rules. In either case, we assume that $\Gamma, x : \pi \vdash \tau[]^l$ checks s , meaning we can use our extended environment lemmas and the **T-Seq** and **T-Free** rules to check these cases.

- Case **T-Alloc-Shared**

Same as previous case.

- Case **T-Partition**

In this case we have by assumption that $\Gamma, y : (h, n/c) \vdash \tau[]^l$'s and l is not local and c divides n . By inversion on our step relation we must be in the **S-Partition** case, so we have $\eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash \pi \text{ partition } \psi \text{ y c s } \sim \text{init}_\psi; s'; \text{dec}_\psi; \text{wait}_\psi \dashv_m \eta', \sigma', \Sigma', \Psi, \Phi$ where $s' = s[(y + c \cdot p)/y]$ and $(\eta', \sigma', \Sigma') = \text{rename}(\eta, \sigma, \Sigma, x, y, \pi/c)$. Via our well-typed renaming lemma contexts we know that we can check the renamed environments in the extended environment $\Gamma, y : \pi/c \vdash \tau[]^l, x : \pi$

$\tau[]^l$, and this context is also sufficient to check s' (via a substitution-preserves-typing lemma that is obvious). Using the **T-Skip** rule this is exactly what we need to show to complete this case, as the thread sync primitives check trivially via their typing rules.

- Case **T-Claim**

Essentially the same as **T-Partition**.

- Case **T-Lower**

Essentially the same as **T-Partition**.

Control Rules.

- Case **T-If**

We have by assumption that **if** e **then** s_1 **else** s_2 steps to some s' , and by inversion we know that either e evaluates to **true** and s' is s_1 , or e evaluates to **false** and s' is s_2 .

In either case, our inductive hypotheses is sufficient to tell us that these are well-typed. In particular, in both cases our IHs tell us that the amount of memory used by stepping each branch of the **if** is less than the amount of memory computed by the type system for each branch. Because the whole **if** expression checks using the greater of the memory usage of m_1 or m_2 (i.e., the memory usage on each branch), the resulting usage for the whole conditional is also bounded by the type system.

- Case **T-Skip**

Trivial: **skip** does not step

- Case **T-Seq**

In this case we have that s_1 and s_2 are both well-typed (with m_1 memory and m_2 memory respectively), and $m = \mathbf{max}(m_1, m_2)$. We also have by inversion that s_1 either steps to **skip** or s'_1 , and our inductive hypothesis tells us that s'_1 is well-typed.

In the former case we can use the **S-Seq-Done** rule to trivially finish the case. In the latter, our IH allows us to finish the case, since m_1 is always $\leq \mathbf{max}(m_1, m_2)$

- Case **T-While**

By inversion, we have that

$$\begin{aligned} & \eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi \mathbf{while } e \mathbf{ do } s \\ & \rightsquigarrow \mathbf{if } e \mathbf{ then } (s; \mathbf{while } e \mathbf{ do } s) \mathbf{ else skip } \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi. \end{aligned}$$

We also have by assumption that e and s are well-typed. With this information, through use of the **T-If**, **T-Seq**, **T-While**, and **T-Skip** rules, we can conclude that the result of this rule is also well-typed.

- Case **T-Function-Call**

By inversion we have that

$$\begin{aligned} & \eta, \sigma, \Sigma, t, b, p, \Psi, \Phi \vdash_m^\pi f(e_1, \dots, e_n) \\ & \rightsquigarrow \mathbf{call } s \mathbf{ with } (x_i : \tau_i \mapsto v_i) @ \{\eta, \sigma, \Sigma, m'\} \dashv_m \eta, \sigma, \Sigma, \Psi, \Phi, \end{aligned}$$

and also that $\Sigma(f) = \{[x_1 : \tau_1, \dots, x_n : \tau_n], s\}$, and $\eta, \sigma, \Sigma, t, b, p \vdash^\pi e_i \Downarrow v_i$, and $m' \leq m$.

We also have from the premises of our case that $f : \pi \mathbf{Fun}(x_1 : \tau_1, \dots, x_n : \tau_n, \pi, m') \in \Gamma$, and $\Gamma \vdash^\pi e_i : \tau_i$, and $m' \leq m$.

Our lemma for expression well-typedness tells us that that each v_i is a well-typed value, and our assumption that $\Gamma \vdash \eta, \sigma, \Sigma$ tells us that $\Sigma(f)$ is a well-typed function and thus that $\mathbf{fn}s \Gamma, x_i : \tau_i \vdash_{m'}^\pi s$. We can take the union of this with Γ to produce $\Gamma, x_i : \tau_i \vdash_{m'}^\pi s$, which clearly checks s and the output environments.

□

B Prism Programs

B.1 Complete TF32 Matrix Multiplication

```

22  with claim(C, scope=block[1], offset= warp_row * N + warp_col) as C_blk:
23      with group(block[1]):
24          A_smem : shared(float[16 * 8]) @ block[1]
25          B_smem : shared(float[8 * 8]) @ block[1]
26          C_smem : shared(float[16 * 8]) @ block[1]
27
28
29      idx : int @ thread[1] = id() * 4
30      with partition(C_smem, dimension=thread[1], offset=idx) as C_thrd:
31          for i in range(0, 4, 1):
32              with group(thread[1]):
33                  C_thrd[i] = 0
34
35      for i in range(0, K_tiles, 1):
36          a_idx : int @ thread[1] = id() * 4
37          # Sync point
38          for j in range(0, 4, 1):
39              flat_idx : int @ thread[1] = a_idx + j
40              row : int @ thread[1] = flat_idx / MMA_K
41              col : int @ thread[1] = flat_idx % MMA_K
42              global_row : int @ thread[1] = warp_row + row
43              global_col : int @ thread[1] = i * MMA_K + col
44              with partition(A_smem, dimension=thread[1], offset= row * MMA_K + col) as A_smem_thrd:
45                  with group(thread[1]):
46                      A_smem_thrd[0] = A[global_row * K + global_col]
47
48          b_idx : int @ thread[1] = id() * 2
49          # Sync point
50          for j in range(0, 2, 1):
51              flat_idx_b : int @ thread[1] = b_idx + j
52              row_b : int @ thread[1] = flat_idx_b / MMA_K
53              col_b : int @ thread[1] = flat_idx_b % MMA_K
54              global_row_b : int @ thread[1] = i * MMA_K + row_b
55              global_col_b : int @ thread[1] = warp_col + col_b
56              with partition(B_smem, dimension=thread[1], offset= row_b * MMA_K + col_b) as B_smem_thrd:
57                  with group(thread[1]):
58                      B_smem_thrd[0] = B[global_row_b * N + global_col_b]
59
60          # Sync point
61          with claim(C_smem, scope=thread[32], offset=0) as C_smem_warp:
62              match split(thread):
63                  case 32:
64                      simple_mma(A_smem, B_smem, C_smem_warp)
65
66      for j in range(0, 4, 1):
67          flat_idx_c : int @ thread[1] = id() * 4 + j
68          row_c : int @ thread[1] = flat_idx_c / MMA_K
69          col_c : int @ thread[1] = flat_idx_c % MMA_K
70          with partition(C_blk, dimension=thread[1], offset= row_c * N + col_c) as C_thrd:
71              with group(thread[1]):
72                  C_thrd[0] = C_smem[row_c * MMA_N + col_c]
73
return

```

B.2 Different Matrix Multiplications Variants

```

1 @prism("global")
2 @requires(grid[], block[], thread[], smem=49000)
3 def my_sgemm_kernel_1( M : int @ grid[],
4                       N : int @ grid[],
5                       K : int @ grid[],
6                       alpha : float @ grid[],
7                       A: ptr(const(float)) @ grid[],
8                       B: ptr(const(float)) @ grid[],
9                       beta : float @ grid[],
10                      C: ptr(float) @ grid[],
11                      block_size : int @ grid[]):
12     with group(grid[]):
13         bid : int @ block[] = id()
14         with partition(C, dimension=block[], offset=0) as C_blk:
15             with group(block[]):
16                 tid : int @ thread[] = bid * block_size + id()
17                 with partition(C_blk, dimension=thread[], offset=0) as C_thrd:
18                     x : const(int) @ thread[] = tid / N
19                     y : const(int) @ thread[] = tid % N
20                     with group(thread[]):
21                         tmp : float @ thread[] = 0
22                         for i in range(0, K, 1):
23                             tmp += A[x * K + i] * B[i * N + y]
24                         C_thrd[x * N + y] = alpha * tmp + beta * C[x * N + y]
25
    return

```

```

1 @prism("global")
2 @requires(grid[], block[], thread[], smem=49000)
3 def my_sgemm_kernel_2( M : int @ grid[],
4                       N : int @ grid[],
5                       K : int @ grid[],
6                       alpha : float @ grid[],
7                       A: ptr(const(float)) @ grid[],
8                       B: ptr(const(float)) @ grid[],
9                       beta : float @ grid[],
10                      C: ptr(float) @ grid[],
11                      tile_size : int @ grid[]):
12     with group(grid[]):
13         global_tid: int @ thread[] = id()
14         total_elements : const(int) @ grid[] = M * N
15         num_tiles_per_row : const(int) @ grid[] = (N + tile_size - 1) / tile_size
16         tile_id : const(int) @ thread[] = global_tid / (tile_size * tile_size)
17         tile_row : const(int) @ thread[] = tile_id / num_tiles_per_row
18         tile_col : const(int) @ thread[] = tile_id % num_tiles_per_row
19         local_id : const(int) @ thread[] = global_tid % (tile_size * tile_size)
20         local_row : const(int) @ thread[] = local_id / tile_size
21         local_col : const(int) @ thread[] = local_id % tile_size
22         cRow : const(int) @ thread[] = tile_row * tile_size + local_row
23         cCol : const(int) @ thread[] = tile_col * tile_size + local_col
24         with partition(C, dimension=thread[], offset=cRow * N + cCol) as C_thrd:
25             with group(thread[]):
26                 if global_tid < total_elements:
27                     if cRow < M and cCol < N:
28                         tmp : float @ thread[] = 0
29                         for i in range(0, K, 1):
30                             tmp += A[cRow * K + i] * B[i * N + cCol]
31                         C_thrd[0] = alpha * tmp + beta * C_thrd[0]
32
    return

```

```

1 @prism("global")
2 @attr("__launch_bounds__((BM * BN) / (TM * TN), 1)")
3 @template([("BM", c_int),
4           ("BN", c_int),
5           ("BK", c_int),
6           ("TM", c_int),
7           ("TN", c_int)])
8 @requires(grid[1], block[1], thread[1], smem=49000)
9 def my_sgemm_kernel_3c(M : int @ grid[1],
10                      N : int @ grid[1],
11                      K : int @ grid[1],
12                      alpha : float @ grid[1],
13                      A: ptr(const(float)) @ grid[1],
14                      B: ptr(const(float)) @ grid[1],
15                      beta : float @ grid[1],
16                      C: ptr(float) @ grid[1],
17                      num_blocks_N : int @ grid[1]):
18     with group(grid[1]):
19         cRow : const(int) @ block[1] = id() / num_blocks_N
20         cCol : const(int) @ block[1] = id() % num_blocks_N
21         totalResultsBlocktile: const(uint) @ grid[1] = BM * BN
22         numThreadsBlocktile: const(uint) @ grid[1] = totalResultsBlocktile / (TM * TN)
23         with partition(C, dimension=block[1], offset = cRow * BM * N + cCol * BN) as C_blk:
24             with partition(A, dimension=block[1], offset = cRow * BM * K) as A_blk:
25                 with partition(B, dimension=block[1], offset = cCol * BN) as B_blk:
26                     with group(block[1]):
27                         As : shared(float[128 * 8]) @ block[1]
28                         Bs : shared(float[128 * 8]) @ block[1]
29                         threadRow : const(int) @ thread[1] = id() / (BN / TN)
30                         threadCol : const(int) @ thread[1] = id() % (BN / TN)
31
32                         innerRowA : const(int) @ thread[1] = id() / BK
33                         innerColA : const(int) @ thread[1] = id() % BK
34                         strideA : const(int) @ thread[1] = numThreadsBlocktile / BN
35
36                         innerRowB : const(int) @ thread[1] = id() / BN
37                         innerColB : const(int) @ thread[1] = id() % BN
38                         strideB : const(int) @ thread[1] = numThreadsBlocktile / BN
39
40                         threadResults : float[64] @ thread[1] = { 0 }
41                         regM : float[8] @ thread[1] = { 0 } # Need to go add the {}
42                         regN : float[8] @ thread[1] = { 0 }
43
44                         i : int @ thread[1] = 0
45
46                         for bkIdx in range(0, K, BK):
47
48                             with partition(As, dimension=thread[1], offset = 0) as As_thrd:
49                                 with partition(Bs, dimension=thread[1], offset = 0) as Bs_thrd:
50                                     with group(thread[1]):
51                                         for loadOffset in range(0, BM, strideA):
52                                             As_thrd[(innerRowA + loadOffset) * BK + innerColA] = \
53                                                 A_blk[i * BK + (innerRowA + loadOffset) * K + innerColA]
54
55                                         for loadOffset in range(0, BK, strideB):
56                                             Bs_thrd[(innerRowB + loadOffset) * BN + innerColB] = \
57                                                 B_blk[(i * BK * N) + (innerRowB + loadOffset) * N + innerColB]
58
59                                         i += 1
60
61                                         for dotIdx in range(0, BK, 1):
62                                             for i in range(0, TN, 1):
63                                                 regM[i] = As[(threadRow * TM + i) * BK + dotIdx]
64                                             for i in range(0, TN, 1):
65                                                 regN[i] = Bs[dotIdx * BN + threadCol * TN + i]
66                                             for resIdxM in range(0, TM, 1):
67                                                 for resIdxN in range(0, TN, 1):
68                                                     threadResults[resIdxM * TN + resIdxN] += regM[resIdxM] * regN[resIdxN]
69
70                                         for resIdxM in range(0, TM, 1):
71                                             for resIdxN in range(0, TN, 1):
72                                                 with partition(C_blk, dimension=thread[1], offset = (threadRow * TM + resIdxM) * N + threadCol * TN +
73                                     resIdxN) as C_thrd:
74                                                     with group(thread[1]):
75                                                         C_thrd[0] = alpha * threadResults[resIdxM * TN + resIdxN] + beta * C_blk[(threadRow * TM + resIdxM) *
76                                     N + threadCol * TN + resIdxN]
77
78     return

```

```

1 @prism("global")
2 @attr("__launch_bounds__((BM * BN) / (TM * TN), 1)")
3 @template([("BM", c_int),
4           ("BN", c_int),
5           ("BK", c_int),
6           ("TM", c_int),
7           ("TN", c_int)])
8 @requires(grid[1], block[1], thread[1], smem=49000)
9 def my_sgemm_kernel_4(M : int @ grid[1],
10                      N : int @ grid[1],
11                      K : int @ grid[1],
12                      alpha : float @ grid[1],
13                      A: ptr[float] @ grid[1],
14                      B: ptr[float] @ grid[1],
15                      beta : float @ grid[1],
16                      C: ptr[float] @ grid[1],
17                      num_blocks_N : int @ grid[1]):
18     with group(grid[1]):
19         cRow : const(int) @ block[1] = id() / num_blocks_N
20         cCol : const(int) @ block[1] = id() % num_blocks_N
21         totalResultsBlocktile: const(uint) @ grid[1] = BM * BN
22         numThreadsBlocktile: const(uint) @ grid[1] = totalResultsBlocktile / (TM * TN)
23         with partition(C, dimension=block[1], offset = cRow * BM * N + cCol * BN) as C_blk:
24             with partition(A, dimension=block[1], offset = cRow * BM * K) as A_blk:
25                 with partition(B, dimension=block[1], offset = cCol * BN) as B_blk:
26                     with group(block[1]):
27                         As : shared(float[128 * 8]) @ block[1]
28                         Bs : shared(float[128 * 8]) @ block[1]
29                         threadRow : const(int) @ thread[1] = id() / (BN / TN)
30                         threadCol : const(int) @ thread[1] = id() % (BN / TN)
31
32                         innerRowA : const(int) @ thread[1] = id() / (BK / 4)
33                         innerColA : const(int) @ thread[1] = id() % (BK / 4)
34                         strideA : const(int) @ thread[1] = numThreadsBlocktile / BK
35
36                         innerRowB : const(int) @ thread[1] = id() / (BN / 4)
37                         innerColB : const(int) @ thread[1] = id() % (BN / 4)
38                         strideB : const(int) @ thread[1] = numThreadsBlocktile / BN
39
40                         threadResults : float[64] @ thread[1] = { 0 }
41                         regM : float[8] @ thread[1] = { 0 } # Need to go add the {}
42                         regN : float[8] @ thread[1] = { 0 }
43
44                         i : int @ thread[1] = 0
45
46                         for bkIdx in range(0, K, BK):
47
48                             with partition(As, dimension=thread[1], offset = 0) as As_thrd:
49                                 with partition(Bs, dimension=thread[1], offset = 0) as Bs_thrd:
50                                     with group(thread[1]):
51
52                                         tmp : float4 @ thread[1] = float4_cast(A_blk[i * BK + innerRowA * K + innerColA * 4])
53                                         As_thrd[(innerColA * 4 + 0) * BM + innerRowA] = tmp.x
54                                         As_thrd[(innerColA * 4 + 1) * BM + innerRowA] = tmp.y
55                                         As_thrd[(innerColA * 4 + 2) * BM + innerRowA] = tmp.z
56                                         As_thrd[(innerColA * 4 + 3) * BM + innerRowA] = tmp.w
57
58                                         i += 1
59
60                                         for dotIdx in range(0, BK, 1):
61                                             for i in range(0, TN, 1):
62                                                 regM[i] = As[dotIdx * BM + threadRow * TM + i]
63                                             for i in range(0, TN, 1):
64                                                 regN[i] = Bs[dotIdx * BN + threadCol * TN + i]
65                                             for resIdxM in range(0, TM, 1):
66                                                 for resIdxN in range(0, TN, 1):
67                                                     threadResults[resIdxM * TN + resIdxN] += regM[resIdxM] * regN[resIdxN]
68
69                                         for resIdxM in range(0, TM, 1):
70                                             for resIdxN in range(0, TN, 4):
71                                                 with partition(C_blk, dimension=thread[1], offset = (threadRow * TM + resIdxM) * N + threadCol * TN +
72                                     resIdxN) as C_thrd:
73                                     with group(thread[1]):
74                                         temp : float4 = float4_cast(C_thrd[0])
75                                         temp.x = alpha * threadResults[resIdxM * TN + resIdxN] + beta * temp.x
76                                         temp.y = alpha * threadResults[resIdxM * TN + resIdxN + 1] + beta * temp.y
77                                         temp.z = alpha * threadResults[resIdxM * TN + resIdxN + 2] + beta * temp.z
78                                         temp.w = alpha * threadResults[resIdxM * TN + resIdxN + 3] + beta * temp.w
79                                         float4_cast[C_thrd[0]] = temp
80
81     return

```

```

1 @prism("global")
2 @attr("__launch_bounds__(NUM_THREADS)")
3 @template([("BM", c_int),
4     ("BN", c_int),
5     ("BK", c_int),
6     ("WM", c_int),
7     ("WN", c_int),
8     ("WNITER", c_int),
9     ("TM", c_int),
10    ("TN", c_int),
11    ("NUM_THREADS", c_int),
12 ])
13 @requires(grid[1], block[], warp[], smem=49000)
14 def my_sgemm_kernel_5(M : int @ grid[1],
15     N : int @ grid[1],
16     K : int @ grid[1],
17     alpha : float @ grid[1],
18     A: ptr(float) @ grid[1],
19     B: ptr(float) @ grid[1],
20     beta : float @ grid[1],
21     C: ptr(float) @ grid[1],
22     num_blocks_N : int @ grid[1]):
23
24     with group(grid[1]):
25         cRow : const(int) @ block[1] = id() / num_blocks_N
26         cCol : const(int) @ block[1] = id() % num_blocks_N
27
28         WMITER : constexpr(uint) @ grid[1] = (WM * WN) / (WARPSPACE * TM * TN * WNITER)
29         WSUBM : constexpr(uint) @ grid[1] = WM / WMITER
30         WSUBN : constexpr(uint) @ grid[1] = WN / WNITER
31
32         rowStrideA : constexpr(int) @ grid[1] = (NUM_THREADS * 4) / BK
33         rowStrideB : constexpr(int) @ grid[1] = NUM_THREADS / (BN / 4)
34         with partition(C, dimension=block[1], offset = cRow * BM * N + cCol * BN) as C_blk:
35             with partition(A, dimension=block[1], offset = cRow * BM * K) as A_blk:
36                 with partition(B, dimension=block[1], offset = cCol * BN) as B_blk:
37                     with group(block[]):
38                         warpId : const(uint) @ thread[32] = id()
39                         warpCol : const(uint) @ thread[32] = warpIdx % (BN / WN)
40                         warpRow : const(uint) @ thread[32] = warpIdx / (BN / WN)
41                         As : shared[float[28 * 16]] @ block[1]
42                         Bs : shared[float[28 * 16]] @ block[1]
43
44                         innerRowA : const(int) @ thread[1] = id() / (BK / 4)
45                         innerColA : const(int) @ thread[1] = id() % (BK / 4)
46
47                         innerRowB : const(int) @ thread[1] = id() / (BN / 4)
48                         innerColB : const(int) @ thread[1] = id() % (BN / 4)
49
50                         threadResults : float[128] @ thread[1] = { 0 }
51                         regM : float[8] @ thread[1] = { 0 } # Need to go add the {}
52                         regN : float[16] @ thread[1] = { 0 }
53                         j : int @ block[1] = 0
54                         for bkIdx in range(0, K, BK):
55                             with partition(A_blk, dimension=block[1], offset = j * BK) as A_Offset:
56                                 with partition(B_blk, dimension=block[1], offset = j * BK * N) as B_Offset:
57                                     loadFromGmem_our(N, K, A_Offset, B_Offset, As, Bs, template=[BM, BN, BK, rowStrideA, rowStrideB])
58                                     processFromSmem_our(regM, regN, threadResults, As, Bs, warpRow, warpCol, template=[BM, BN, BK, WM, WN,
59                                     WMITER, WNITER, WSUBM, WSUBN, TM, TN])
60                                     j+=1
61                                     pass
62
63                         with partition(C_blk, dimension=warp[1], offset = warpRow * WM * N + warpCol * WN) as C_warp:
64                             with group(warp[1]):
65
66                                 threadIdxInWarp : int @ thread[1] = id()
67                                 threadColInWarp : const(uint) @ thread[1] = threadIdxInWarp % (WSUBN / TN)
68                                 threadRowInWarp : const(uint) @ thread[1] = threadIdxInWarp / (WSUBN / TN)
69
70                                 for wSubRowIdx in range(0, WMITER, 1):
71                                     for wSubColIdx in range(0, WNITER, 1):
72                                         with partition(C_warp, dimension=thread[1], offset = (wSubRowIdx * WSUBM) * N + wSubColIdx * WSUBN) as C_interim:
73                                             for resIdxM in range(0, TM, 1):
74                                                 for resIdxN in range(0, TN, 4):
75                                                     with group(thread[1]):
76                                                         tmp : float4 @ thread[1] = float4_cast(C_interim[threadRowInWarp * TM + resIdxM] * N +
77                                                         threadColInWarp * TN + resIdxN)
78                                                         i : const(int) @ thread[1] = (wSubRowIdx * TM + resIdxM) * (WNITER * TN) + wSubColIdx * TN
79                                                         tmp.x = alpha * threadResults[i + 0] + beta * tmp.x
80                                                         tmp.y = alpha * threadResults[i + 1] + beta * tmp.y
81                                                         tmp.z = alpha * threadResults[i + 2] + beta * tmp.z
82                                                         tmp.w = alpha * threadResults[i + 3] + beta * tmp.w
83                                                         float4_cast(C_interim[threadRowInWarp * TM + resIdxM] * N +
84                                                         threadColInWarp * TN + resIdxN] = tmp
85
86             return

```

B.3 Single-Pass Parallel Prefix Scan with Decoupled Look-Back

```

1 @prism("global")
2 @requires(grid[], thread[])
3 def add_m_thread(x: ptr(const(c_int)) @ grid[], 
4                  y: ptr(c_int) @ grid[], 
5                  m: uint32_t @ grid[]):
6     with group(grid[]):
7         tid : uint32_t @ thread[] = id()
8         with partition(x, offset=tid, dimension=thread[]) as x_thread:
9             with partition(y, offset=tid, dimension=thread[]) as y_thread:
10                with group(thread[]):
11                    y_thread[0] = x_thread[0] + m
12    return
13
14
15 # WORK_PER_BLOCK = (8192 * 2)
16 # WARPS_PER_BLOCK = 16
17 # WARPS_PASS_CNT = (8192 * 2) / (16 * 32) = 32
18 # WARP_ENDS = (8192 * 2) / 32 = 512
19
20 @prism("device")
21 @requires(thread[])
22 def shfl_up_sync_wrapper(val : uint32_t @ thread[], delta : uint32_t @ thread[]):
23     pass
24
25 @prism("device")
26 @requires(block[], warp[16])
27 def scan_warp(x : ptr(uint32_t) @ block[], x_shmem : ptr(uint32_t) @ block[]):
28     with group(block[]):
29         for i in range(0, (8192 * 2) / (16 * 32)):
30             warp_id : uint32_t @ warp[] = id()
31             with partition(x, offset = (warp_id + i * 16) * 32, dimension = warp[1]) as x_warp:
32                 with partition(x_shmem, offset = (warp_id + i * 16) * 32, dimension = warp[1]) as x_shmem_warp:
33                     with group(warp[]):
34                         thread_id : uint32_t @ thread[] = id()
35                         with partition(x_warp, offset = thread_id, dimension = thread[]) as x_thread:
36                             with partition(x_shmem_warp, offset = thread_id, dimension = thread[]) as x_shmem_thread:
37                                 with group(thread[]):
38                                     val : uint32_t @ thread[] = x_thread[0]
39
40                                     stride : uint32_t @ thread[] = 1
41                                     while stride < 32:
42                                         received : uint32_t @ thread[] = shfl_up_sync_wrapper(val, stride)
43                                         if thread_id >= stride:
44                                             val = received + val
45                                         stride *= 2
46                                         x_shmem_thread[0] = val
47
48 @prism("device")
49 @requires(block[], warp[16])
50 def get_warp_ends_inclusive(x : ptr(uint32_t) @ block[], warp_ends : ptr(uint32_t) @ block[]):
51     with group(block[]):
52         warp_id : uint32_t @ warp[] = id()
53         with partition(x, offset = (warp_id) * 32 * 32, dimension = warp[1]) as x_warp:
54             with partition(warp_ends, offset = (warp_id) * 32, dimension = warp[1]) as warp_ends_warp:
55                 with group(warp[]):
56                     thread_id : uint32_t @ thread[] = id()
57                     with partition(x_warp, offset = thread_id * 32, dimension = thread[]) as x_thread:
58                         with partition(warp_ends_warp, offset = thread_id, dimension = thread[]) as warp_ends_thread:
59                             with group(thread[]):
60                                 warp_ends_thread[0] = x_thread[31]
61
62
63
64 @prism("device")
65 @requires(block[], warp[16])
66 def update_with_new_data_and_writeback(x : ptr(uint32_t) @ block[], x_shmem : ptr(uint32_t) @ block[], warp_ends : ptr(uint32_t) @
67 block[], prefix : uint32_t @ block[]):
68     with group(block[]):
69         for i in range(0, (8192 * 2) / (16 * 32)):
70             warp_id : uint32_t @ warp[] = id()
71             with partition(x, offset = (warp_id + i * 16) * 32, dimension = warp[1]) as x_warp:
72                 with partition(x_shmem, offset = (warp_id + i * 16) * 32, dimension = warp[1]) as x_shmem_warp:
73                     with group(warp[]):
74                         thread_id : uint32_t @ thread[] = id()
75                         with partition(x_warp, offset = thread_id, dimension = thread[]) as x_thread:
76                             with partition(x_shmem_warp, offset = thread_id, dimension = thread[]) as x_shmem_thread:
77                                 with group(thread[]):
78                                     if warp_id + i * 16 == 0:
79                                         x_thread[0] = prefix + x_shmem_thread[0]

```

```

79             else:
80                 x_thread[0] = prefix + warp_ends[warp_id + i * 16 - 1] + x_shmem_thread[0]
81
82     @prism("device")
83     @requires(block[1], warp[16])
84     def naive_scan_warp_ends(x : ptr(uint32_t) @ block[1], x_buf : ptr(uint32_t) @ block[1]):
85         with group(block[1]):
86             stride : uint32_t @ group[1] = 1
87             while stride < (8192 * 2 / 32):
88
89                 warp_id : uint32_t @ warp[1] = id()
90                 with partition(x, offset = warp_id * 32, dimension = warp[1]) as x_warp:
91                     with partition(x_buf, offset = warp_id * 32, dimension = warp[1]) as x_buf_warp:
92                         with group(warp[1]):
93                             thread_id : uint32_t @ thread[1] = id()
94                             with partition(x_warp, offset = thread_id, dimension = thread[1]) as x_thread:
95                                 with partition(x_buf_warp, offset = thread_id, dimension = thread[1]) as x_buf_thread:
96                                     with group(thread[1]):
97                                         if thread_id + warp_id * 32 >= stride:
98                                             x_buf_thread[0] = x_thread[0] + x_thread[-stride]
99                                         else:
100                                             x_buf_thread[0] = x_thread[0]
101
102                 tmp : ptr(uint32_t) @ block[1] = x
103                 x = x_buf
104                 x_buf = tmp
105
106                 stride *= 2
107             return x
108
109
110 # Implemented in raw CUDA: only exposing the necessary interfaces
111 @prism("device")
112 @requires(block[1], thread[1])
113 def write_block_aggregate(aggregate : uint32_t @ block[1], block_idx : uint32_t @ block[1], block_infos : ptr(int) @ block[1]):
114     pass
115
116 @prism("device")
117 @requires(block[1], thread[1])
118 def write_block_prefix(prefix : uint32_t @ block[1], block_idx : uint32_t @ block[1], block_infos : ptr(int) @ block[1]):
119     pass
120
121
122 @prism("device")
123 @requires(block[1], thread[1])
124 def lookback_to_prev_blocks(block_idx : uint32_t @ block[1], block_infos : ptr(int) @ block[1]):
125     pass
126
127 @prism("device")
128 @requires(grid[1])
129 def get_block_idx_from_global_counter(block_idx_counter : ptr(uint32_t) @ grid[1]):
130     pass
131
132 @prism("global")
133 @requires(grid[1], block[1], warp[16], smem=69632)
134 def scan_kernel(n : uint32_t @ grid[1], x : ptr(uint32_t) @ grid[1], block_idx_counter : ptr(uint32_t) @ grid[1], block_infos :
135     → ptr(int) @ grid[1]):
136     block_idx : uint32_t @ grid[1] = get_block_idx_from_global_counter(block_idx_counter)
137
138     with partition(x, offset = (8192 * 2) * block_idx, dimension = block[1]) as x_block:
139         with group(block[1]):
140             x_shmem: shared(uint32_t[8192 * 2]) @ block[1]
141
142             scan_warps(x_block, x_shmem)
143
144             warp_ends : shared(uint32_t[512]) @ block[1]
145             warp_ends_buf : shared(uint32_t[512]) @ block[1]
146
147             get_warp_ends_inclusive(x_shmem, warp_ends)
148             warp_ends_res : ptr(uint32_t) @ block[1] = naive_scan_warp_ends(warp_ends, warp_ends_buf)
149
150             aggregate : uint32_t @ block[1] = warp_ends_res[512 - 1]
151             write_block_aggregate(aggregate, block_idx, block_infos)
152
153             prefix : uint32_t @ block[1] = lookback_to_prev_blocks(block_idx, block_infos)
154             write_block_prefix(aggregate + prefix, block_idx, block_infos)
155
156             update_with_new_data_and_writeback(x_block, x_shmem, warp_ends_res, prefix)

```

B.4 CUB Functions

```

1 @prism("device")
2 @requires(block[1], thread[32])
3 def block_load(input : ptr(const(int)) @ block[1],
4                 output : ptr(int) @ thread[1],
5                 items_per_thread : int @ block[1]):
6     with group(block[1]):
7         tid : int @ thread[32] = id()
8         with partition(input, dimension=thread[32], offset = tid * items_per_thread) as input_thrd:
9             with group(thread[32]):
10                 warp_load(input_thrd, output, items_per_thread)
11
12     return

```

```

1 @prism("device")
2 @requires(thread[32])
3 def warp_load(input : ptr(const(int)) @ thread[32],
4                 output : ptr(int) @ thread[1],
5                 items_per_thread : int @ thread[32]):
6     with group(thread[32]):
7         tid : int @ thread[] = id()
8         with partition(input, dimension=thread[1], offset = tid * items_per_thread) as input_thrd:
9             with group(thread[1]):
10                 thread_load(input_thrd, output, items_per_thread)
11
12     return

```

```

1 @prism("device")
2 @requires(thread[1])
3 def thread_load(input : ptr(const(int)) @ thread[1],
4                 output : ptr(int) @ thread[1],
5                 items_per_thread : int @ thread[1]):
6     with group(thread[1]):
7         for i in range(0, items_per_thread, 1):
8             output[i] = input[i]
9
10    return

```

```

1 @prism("device")
2 @requires(block[1], thread[32])
3 def block_store(input : ptr(const(int)) @ thread[1],
4                  output : ptr(int) @ block[1],
5                  items_per_thread : int @ block[1]):
6     with group(block[1]):
7         tid : int @ thread[32] = id()
8         with partition(output, dimension=thread[32], offset = tid * items_per_thread) as output_thrd:
9             with group(thread[32]):
10                 warp_store(input, output_thrd, items_per_thread)
11
12    return

```

```

1 @prism("device")
2 @requires(thread[32])
3 def warp_store(input : ptr(const(int)) @ thread[1],
4                 output : ptr(int) @ thread[32],
5                 items_per_thread : int @ thread[32]):
6     with group(thread[32]):
7         tid : int @ thread[1] = id()
8         with partition(output, dimension=thread[1], offset = tid * items_per_thread) as output_thrd:
9             with group(thread[1]):
10                 thread_store(input, output_thrd, items_per_thread)
11
12    return

```

```

1 @prism("device")
2 @requires(thread[1])
3 def thread_store(input : ptr(const(int)) @ thread[1],
4                   output : ptr(int) @ thread[1],
5                   items_per_thread : int @ thread[1]):
6     with group(thread[1]):
7         for i in range(0, items_per_thread, 1):
8             output[i] = input[i]
9
10    return

```

B.5 H100 Matrix Multiplication

```

1  @prism("global")
2  @attr("__launch_bounds__(128*3)")
3  @requires(grid[1], block[1], warp[12], smem=227000)
4  def my_h100_mm_no_tail(
5      M: int @ grid[1],
6      N: int @ grid[1],
7      K: int @ grid[1],
8      C: ptr(bf16) @ grid[1],
9      tensorMapA: const(CUtensorMap) @ grid[1],
10     tensorMapB: const(CUtensorMap) @ grid[1]):
11
12     with group(grid[1]):
13         BM : constexpr(int) @ grid[1] = 128
14         BN : constexpr(int) @ grid[1] = 256
15         BK : constexpr(int) @ grid[1] = 64
16         NUM_THREADS : constexpr(int) @ grid[1] = 128*3
17         QSIZE : constexpr(int) @ grid[1] = 4
18         NUM_SM : constexpr(int) @ grid[1] = 128
19
20         WGMMA_M : constexpr(int) @ grid[1] = 64
21         WGMMA_K : constexpr(int) @ grid[1] = 16
22         WGMMA_N : constexpr(int) @ grid[1] = BN
23         num_consumers : constexpr(int) @ grid[1] = (NUM_THREADS / 128) - 1
24         B_WG_M : constexpr(int) @ grid[1] = BM / num_consumers
25
26         TM : constexpr(int) @ grid[1] = 16
27         TN : constexpr(int) @ grid[1] = 8
28         schedule_block : int @ block[1] = id()
29
30     with partition(C, offset=0, dimension=block[1]) as block_C:
31         with group(block[1]):
32
33             # Allocate SA
34             # sA[QSIZE][BK*BM]
35             sA_slot0 : shared(bf16[64* 128], align=128) @ block[1]
36             sA_slot1 : shared(bf16[64* 128]) @ block[1]
37             sA_slot2 : shared(bf16[64* 128]) @ block[1]
38             sA_slot3 : shared(bf16[64* 128]) @ block[1]
39
40             # Allocate SB
41             # sB[QSIZE][BK*BN]
42             sB_slot0 : shared(bf16[64* 256], align=128) @ block[1]
43             sB_slot1 : shared(bf16[64* 256]) @ block[1]
44             sB_slot2 : shared(bf16[64* 256]) @ block[1]
45             sB_slot3 : shared(bf16[64* 256]) @ block[1]
46
47
48             num_blocks_k : const(int) @ block[1] = K / BK
49             wg_idx : int @ warp[4] = id()
50             blk_thrd_id : int @ thread[1] = id()
51
52
53             tid : int @ thread[1] = id() % 128
54             is_producer : bool @ warp[4] = wg_idx == 0
55
56             num_block_m : int @ block[] = 0
57             num_block_n : int @ block[] = 0
58
59             with group(warp[4]):
60                 if (is_producer):
61                     pass
62                 else:
63                     wg_idx = wg_idx - 1
64
65             schedule_it : int @ block[1] = 0
66             total_blocks_m : int @ block[1] = (((M + BM) - 1) / BM)
67             total_blocks_n : int @ block[1] = (((N + BN) - 1) / BN)
68             unsafe("assert(CEIL_DIV(M, BM)%TM == 0 & total_blocks_n%TN == 0);")
69
70             while(true):
71                 num : int @ block[1] = ((schedule_it * NUM_SM) + schedule_block)
72                 if (num >= (total_blocks_m * total_blocks_n)):
73                     break
74                 cur_tile : int @ block[1] = (num / (TM * TN))
75                 cur_tile_pos : int @ block[1] = (num % (TM * TN))
76                 num_block_m = (TM * (cur_tile / (total_blocks_n / TN)))
77                 num_block_n = (TN * (cur_tile % (total_blocks_n / TN)))
78                 num_block_m += (cur_tile_pos / TN)
79                 num_block_n += (cur_tile_pos % TN)

```

```

80     schedule_it += 1
81     d : c_float[1][16][8] @ thread[1]
82
83     with group(thread[1]):
84         d_ptr : ptr(float) @ thread[1]
85         # TODO: ANNOYING TO NOT HAVE TO SET EACH INDIVIDUAL TO 0.
86         unsafe("d_ptr = (float *)d;")
87         memset_wrapper(d_ptr, 0, 1 * 16 * 8 * 4)
88
89     # Now remove these registers...
90     with group(warp[4]):
91         if (is_producer):
92             num_regs : constexpr(int) @ warp[4] = 24 if (num_consumers <= 2) else 32
93             warpgroup_reg_dealloc(template=[num_regs])
94             pass
95         else:
96             num_regs_p : constexpr(int) @ warp[4] = 256 if num_consumers == 1 else (240 if num_consumers == 2 else 160)
97             warpgroup_reg_alloc(template=[num_regs_p])
98             pass
99
100    num_blocks_k_main : const(int) @ block[1] = (num_blocks_k / 4) * 4
101
102    for block_k_iter in range(0, num_blocks_k_main, 4):
103
104        with claim(sA_slot0, scope=thread[1], offset=0) as sA_new:
105            with claim(sA_slot1, scope=thread[1], offset=0) as sA_new1:
106                with claim(sA_slot2, scope=thread[1], offset=0) as sA_new2:
107                    with claim(sB_slot0, scope=thread[1], offset=0) as sB_new:
108                        with claim(sB_slot1, scope=thread[1], offset=0) as sB_new1:
109                            with claim(sB_slot2, scope=thread[1], offset=0) as sB_new2:
110                                with claim(sB_slot3, scope=thread[1], offset=0) as sB_new3:
111                                    with claim(sB_slot0, scope=thread[1], offset=0) as sB_new:
112                                        match warp:
113                                            case 8:
114                                                pass
115                                            case 4:
116                                                match thread:
117                                                    case 1:
118                                                        expect_bytes(full_0, (BK*BN+BK*BM)*2)
119                                                        # TODO: BECOMES AN INTRINSIC CALL!
120                                                        ta_void_ptr : ptr(const(void)) @ thread[1]
121                                                        unsafe("ta_void_ptr = &tensorMapA;")
122                                                        load_async(sA_new, ta_void_ptr, full_0, (block_k_iter+0)*BK, num_block_m*BM)
123                                                        tb_void_ptr : ptr(const(void)) @ thread[1]
124                                                        unsafe("tb_void_ptr = &tensorMapB;")
125                                                        load_async(sB_new, tb_void_ptr, full_0, (block_k_iter+0)*BK, num_block_n*BN)
126
127
128                                                    expect_bytes(full_1, (BK*BN+BK*BM)*2)
129                                                    load_async(sA_new1, ta_void_ptr, full_1, (block_k_iter+1)*BK, num_block_m*BM)
130                                                    load_async(sB_new1, tb_void_ptr, full_1, (block_k_iter+1)*BK, num_block_n*BN)
131
132
133                                                    expect_bytes(full_2, (BK*BN+BK*BM)*2)
134                                                    load_async(sA_new2, ta_void_ptr, full_2, (block_k_iter+2)*BK, num_block_m*BM)
135                                                    load_async(sB_new2, tb_void_ptr, full_2, (block_k_iter+2)*BK, num_block_n*BN)
136
137                                                    expect_bytes(full_3, (BK*BN+BK*BM)*2)
138                                                    load_async(sA_new3, ta_void_ptr, full_3, (block_k_iter+3)*BK, num_block_m*BM)
139                                                    load_async(sB_new3, tb_void_ptr, full_3, (block_k_iter+3)*BK, num_block_n*BN)
140
141        with claim(sA_slot0, scope=warp[8], offset=0) as wmma_sA:
142            with claim(sA_slot1, scope=warp[8], offset=0) as sA_1_producer:
143                with claim(sA_slot2, scope=warp[8], offset=0) as sA_2_producer:
144                    with claim(sB_slot0, scope=warp[8], offset=0) as wmma_sb:
145                        with claim(sB_slot1, scope=warp[8], offset=0) as sB_1_producer:
146                            with claim(sB_slot2, scope=warp[8], offset=0) as sB_2_producer:
147                                with claim(sA_slot3, scope=warp[8], offset=0) as sA_3_producer:
148                                    with claim(sB_slot3, scope=warp[8], offset=0) as sB_3_producer:
149                                    match warp:
150                                        case 8:
151                                            with group(warp[4]):
152                                                with wmma_async():
153                                                    for m_it in range(0, B_WG_M/WGMMA_M, 1):
154                                                        index_a : int @ warp[4] = m_it * BM
155                                                        index_b : int @ warp[4] = m_it * BN
156                                                        unsafe("#pragma unroll")
157                                                        for k_it in range(0, 64/WGMMA_K, 1):
158                                                            intrinsic.wmma256(wmma_sA[64*(m_it + wg_idx*B_WG_M/WGMMA_M)*WGMMA_M +
159                                                                 $\leftrightarrow$  index_a + k_it*WGMMA_K], wmma_sb[(index_b + (k_it *
160                                                                 $\leftrightarrow$  WGMMA_K)), 1, 1, 0, 0, out[d[m_it]]])

```

```

159
160         with wmma_async():
161             for m_it in range(0, B_WG_M/WGMMA_M, 1):
162                 index_a2 : int @ warp[4] = m_it * BM
163                 # TOTALLY OKAY! I just don't want to add support for this the compiler right now
164                 unsafe("#pragma unroll")
165                 for k_it in range(0, 64/WGMMA_K, 1):
166                     intrinsic.wmma256(sA_1_producer[64*(m_it + wg_idx*B_WG_M/WGMMA_M)*WGMMA_M
167                                     ↪ + index_a2 + k_it*WGMMA_K], sB_1_producer[(index_b2 + (k_it *
168                                     ↪ WGMMA_K))], 1, 1, 1, 0, 0, out=[d[m_it]])
169
170
171         with wmma_async():
172             for m_it in range(0, B_WG_M/WGMMA_M, 1):
173                 index_a3 : int @ warp[4] = m_it * BM
174                 index_b3 : int @ warp[4] = m_it * BN
175                 unsafe("#pragma unroll")
176                 for k_it in range(0, 64/WGMMA_K, 1):
177                     intrinsic.wmma256(sA_2_producer[64*(m_it + wg_idx*B_WG_M/WGMMA_M)*WGMMA_M
178                                     ↪ + index_a3 + k_it*WGMMA_K], sB_2_producer[(index_b3 + (k_it *
179                                     ↪ WGMMA_K))], 1, 1, 1, 0, 0, out=[d[m_it]])
180
181         with wmma_async():
182             for m_it in range(0, B_WG_M/WGMMA_M, 1):
183                 index_a4 : int @ warp[4] = m_it * BM
184                 index_b4 : int @ warp[4] = m_it * BN
185                 unsafe("#pragma unroll")
186                 for k_it in range(0, 64/WGMMA_K, 1):
187                     intrinsic.wmma256(sA_3_producer[64*(m_it + wg_idx*B_WG_M/WGMMA_M)*WGMMA_M
188                                     ↪ + index_a4 + k_it*WGMMA_K], sB_3_producer[(index_b4 + (k_it *
189                                     ↪ WGMMA_K))], 1, 1, 1, 0, 0, out=[d[m_it]])
190
191     with claim(block_C, scope=warp[8], offset=num_block_n*BN*M + num_block_m*BM) as c_consumer:
192         match warp:
193             case 8:
194                 with partition(c_consumer, dimension=thread[1], offset=0) as c_thrd:
195                     with group(warp[4]):
196                         warp_id: int @ warp[1] = id()
197                         with group(warp[1]):
198                             lane : int @ thread[1] = id()
199                             row : int @ thread[1] = warp_id * 16 + lane / 4
200                             unsafe("#pragma unroll")
201                             for m_it in range(0, B_WG_M/WGMMA_M, 1):
202                                 yo : int @ warp[1] = m_it*WGMMA_M + wg_idx*B_WG_M
203                                 with group(thread[1]):
204                                     if (row + yo + num_block_m*BM >= M):
205                                         continue
206                                     for w in range(0, WGMMNA_N, 16):
207                                         if (w < w < N-num_block_n*BN):
208                                             col : int @ thread[1] = w + 2*(tid % 4);
209                                             c_thrd[col * M + row + yo] = d[m_it][w/16][0]
210                                             c_thrd[(col + 1) * M + (row + 1) * yo] = d[m_it][w/16][1]
211                                             c_thrd[(col) * M + (row + 8) * yo] = d[m_it][w/16][2]
212                                             c_thrd[(col + 1) * M + (row + 8) * yo] = d[m_it][w/16][3]
213                                             c_thrd[(col + 8) * M + (row + 0) * yo] = d[m_it][w/16][4]
214                                             c_thrd[(col + 9) * M + (row + 0) * yo] = d[m_it][w/16][5]
215                                             c_thrd[(col + 8) * M + (row + 8) * yo] = d[m_it][w/16][6]
216                                             c_thrd[(col + 9) * M + (row + 8) * yo] = d[m_it][w/16][7]
217
218             case 4:
219                 pass
220
221     return

```