

(De/Re)-Composition of Array Computations via Multi-Dimensional Homomorphisms

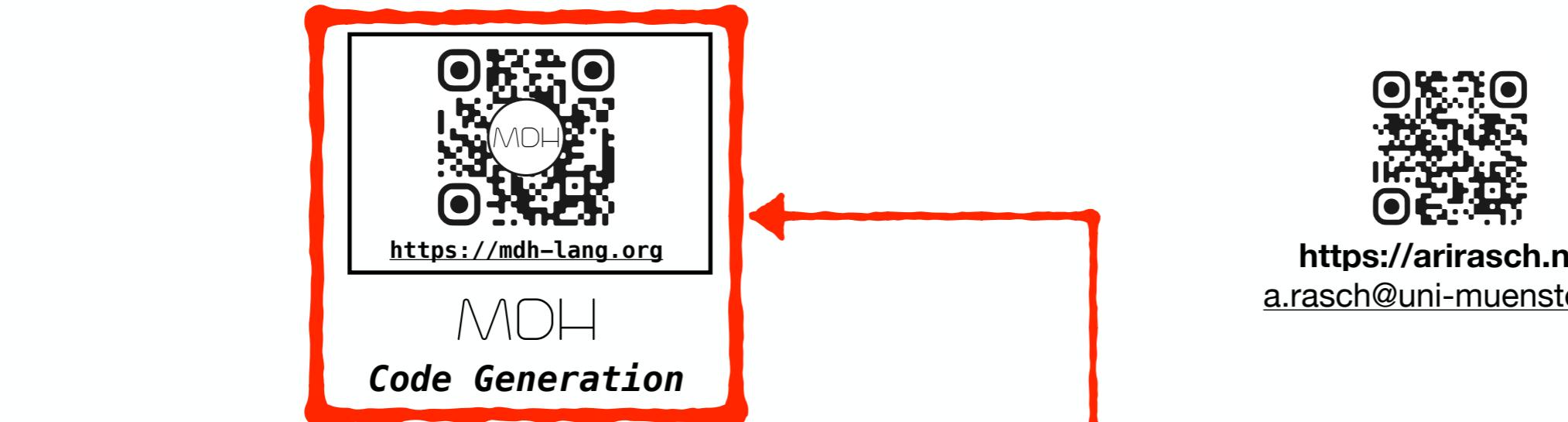
+

An MDH-Based DSL for Array Computations

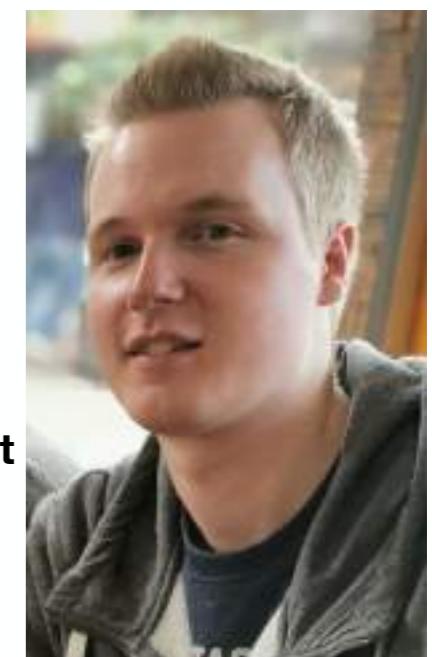
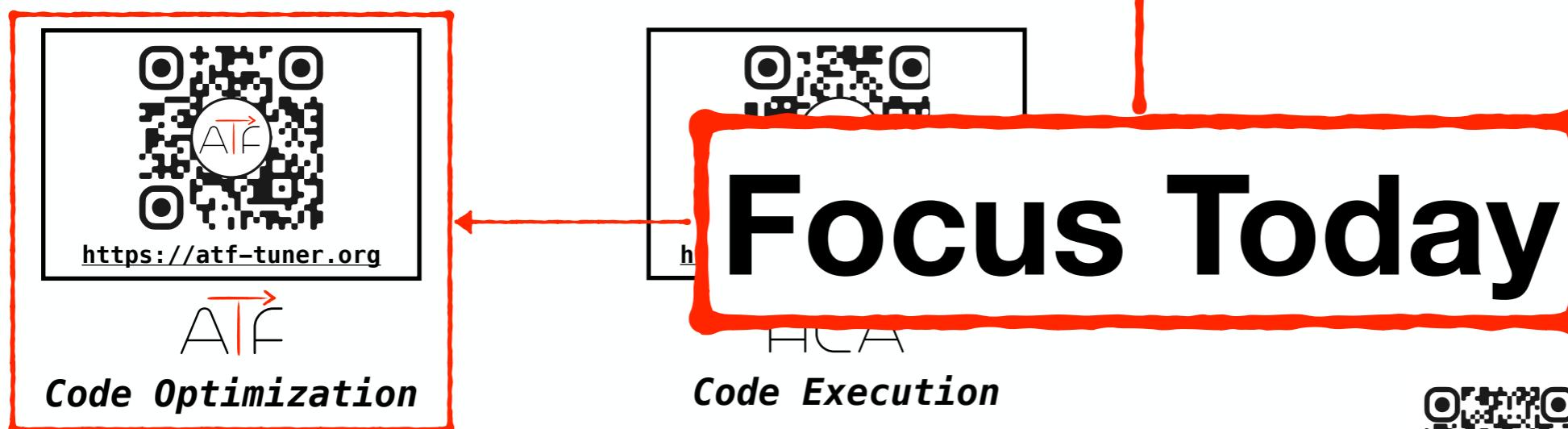
Ari Rasch, Richard Schulze
University of Münster, Germany

Who are we?

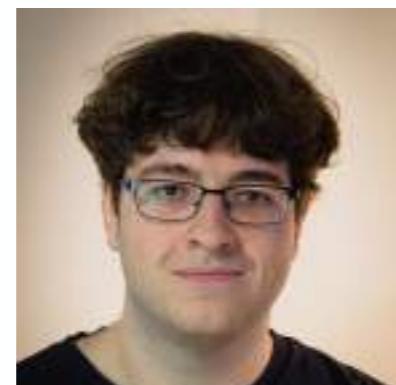
Main Projects:



Ari
Rasch



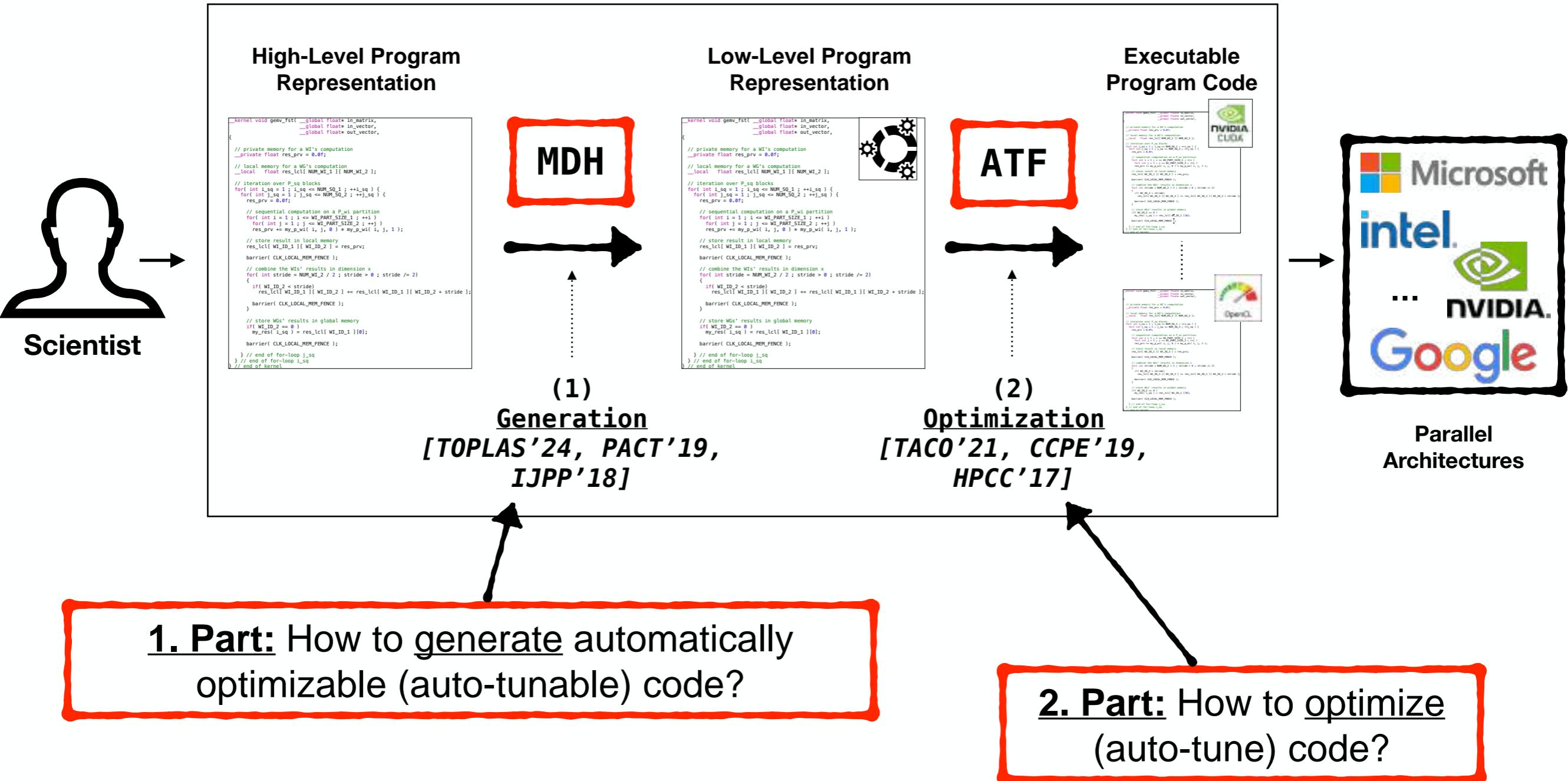
Richard
Schulze



Hunloch Brothers
(Lars & Jens)

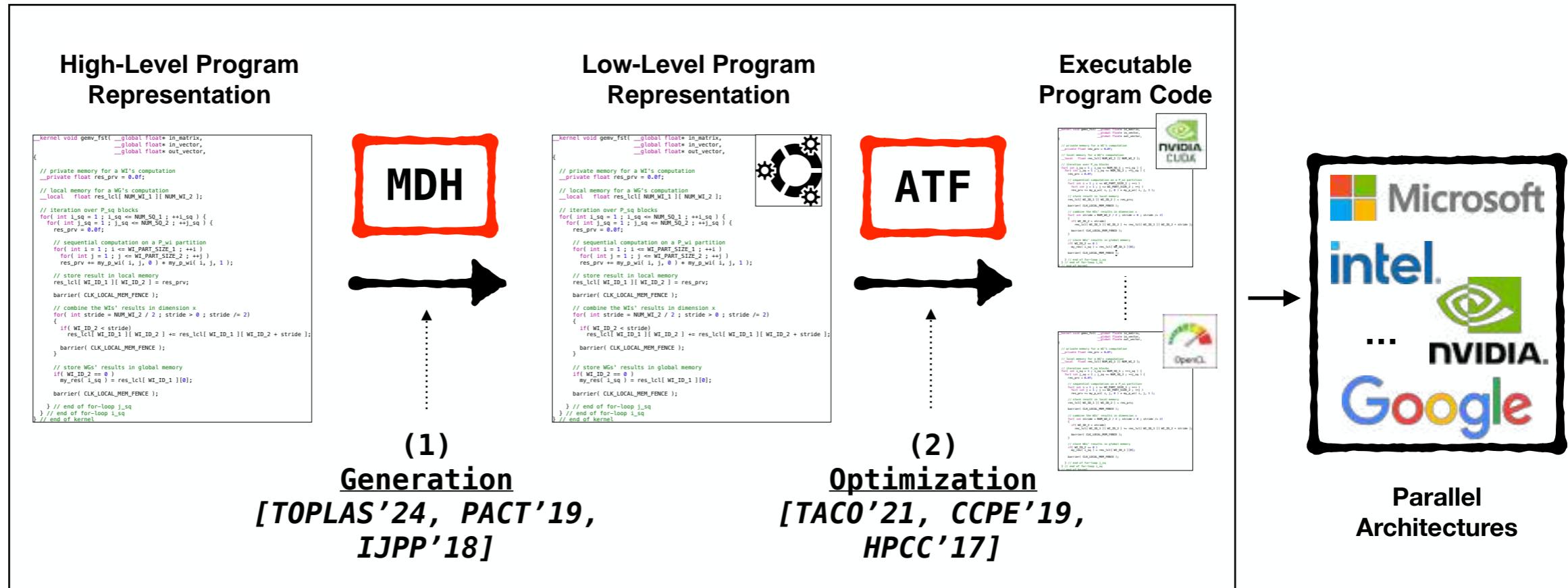
Goal of MDH+ATF

An approach to **Generating (MDH)** & **Optimizing (ATF)** code for **data-parallel computations**:



Goal of MDH+ATF

An approach to **Generating (MDH)** & **Optimizing (ATF)** code for **data-parallel computations**:



The ultimate goal of **MDH+ATF** is to simultaneously achieve
Performance & Portability & Productivity
for data-parallel computations

Code Generation via MDH



Overview Getting Started Code Examples Publications Citations Contact



Multi-Dimensional Homomorphisms (MDH)

An Algebraic Approach Toward Performance & Portability & Productivity
for Data-Parallel Computations

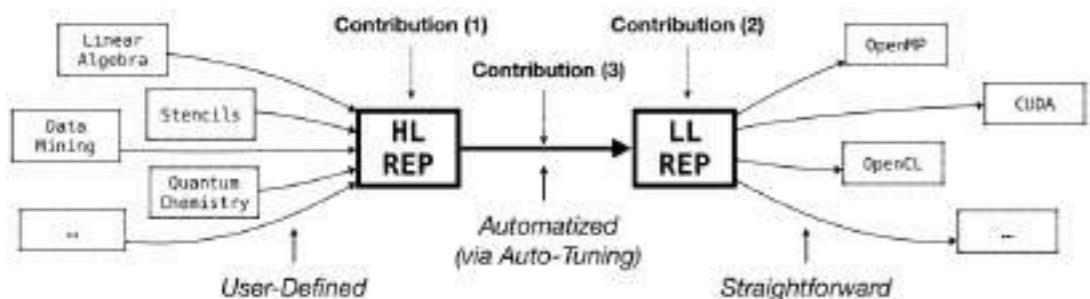
Overview

The approach of **Multi-Dimensional Homomorphisms (MDH)** is an algebraic formalism for systematically reasoning about *de-composition* and *re-composition* strategies of data-parallel computations (such as linear algebra routines and stencil computations) for the memory and core hierarchies of state-of-the-art parallel architectures (GPUs, multi-core CPU, multi-device and multi-node systems, etc).

The MDH approach (formally) introduces:

1. *High-Level Program Representation* (*Contribution 1*) that enables the user conveniently implementing data-parallel computations, agnostic from hardware and optimization details;
2. *Low-Level Program Representation* (*Contribution 2*) that expresses device- and data-optimized de- and re-composition strategies of computations;
3. *Lowering Process* (*Contribution 3*) that fully automatically lowers a data-parallel computation expressed in its high-level program representation to an optimized instance in its low-level representation, based on concepts from automatic performance optimization (a.k.a. *auto-tuning*), using the *Auto-Tuning Framework (ATF)*.

The MDH's low-level representation is designed such that *Code Generation* from it (e.g., in *OpenMP* for CPUs, *CUDA* for NVIDIA GPUs, or *OpenCL* for multiple kinds of architectures) becomes straightforward.



Our *Experiments* report encouraging results on GPUs and CPUs for MDH as compared to state-of-practice approaches, including NVIDIA *cuBLAS/cuDNN* and Intel *oneMKL/oneDNN* which are hand-optimized libraries provided by vendors.

<https://mdh-lang.org>

ACM TOPLAS 2024

(De/Re)-Composition of Data-Parallel Computations via Multi-Dimensional Homomorphisms

ARI RASCH, University of Muenster, Germany

Data-parallel computations, such as linear algebra routines and stencil computations, constitute one of the most relevant classes in parallel computing, e.g., due to their importance for deep learning. Efficiently de-composing such computations for the memory and core hierarchies of modern architectures and re-composing the computed intermediate results back to the final result—we say *(de/re)-composition* for short—is key to achieve high performance for these computations on, e.g., GPU and CPU. Current high-level approaches to generating data-parallel code are often restricted to a particular subclass of data-parallel computations and architectures (e.g., only linear algebra routines on only GPU or only stencil computations), and/or the approaches rely on a user-guided optimization process for a well-performing (de/re)-composition of computations, which is complex and error prone for the user.

We formally introduce a systematic (de/re)-composition approach, based on the algebraic formalism of Multi-Dimensional Homomorphisms (MDHs). Our approach is designed as general enough to be applicable to a wide range of data-parallel computations and for various kinds of target parallel architectures. To efficiently target the deep and complex memory and core hierarchies of contemporary architectures, we exploit our introduced (de/re)-composition approach for a correct-by-construction, parametrized cache blocking, and parallelization strategy. We show that our approach is powerful enough to express, in the same formalism, the (de/re)-composition strategies of different classes of state-of-the-art approaches (scheduling-based, polyhedral, etc.), and we demonstrate that the parameters of our strategies enable systematically generating code that can be fully automatically optimized (auto-tuned) for the particular target architecture and characteristics of the input and output data (e.g., their sizes and memory layouts). Particularly, our experiments confirm that via auto-tuning, we achieve higher performance than state-of-the-art approaches, including hand-optimized solutions provided by vendors (such as NVIDIA cuBLAS/cuDNN and Intel oneMKL/oneDNN), on real-world datasets and for a variety of data-parallel computations, including linear algebra routines, stencil and quantum chemistry computations, data mining algorithms, and computations that recently gained high attention due to their relevance for deep learning.

CCS Concepts: • Computing methodologies → Parallel computing methodologies; Machine learning;
• Theory of computation → Program semantics; • Software and its engineering → Compilers;

Additional Key Words and Phrases: Code generation, data parallelism, auto-tuning, GPU, CPU, OpenMP, CUDA, OpenCL, linear algebra, stencils computation, quantum chemistry, data mining, deep learning

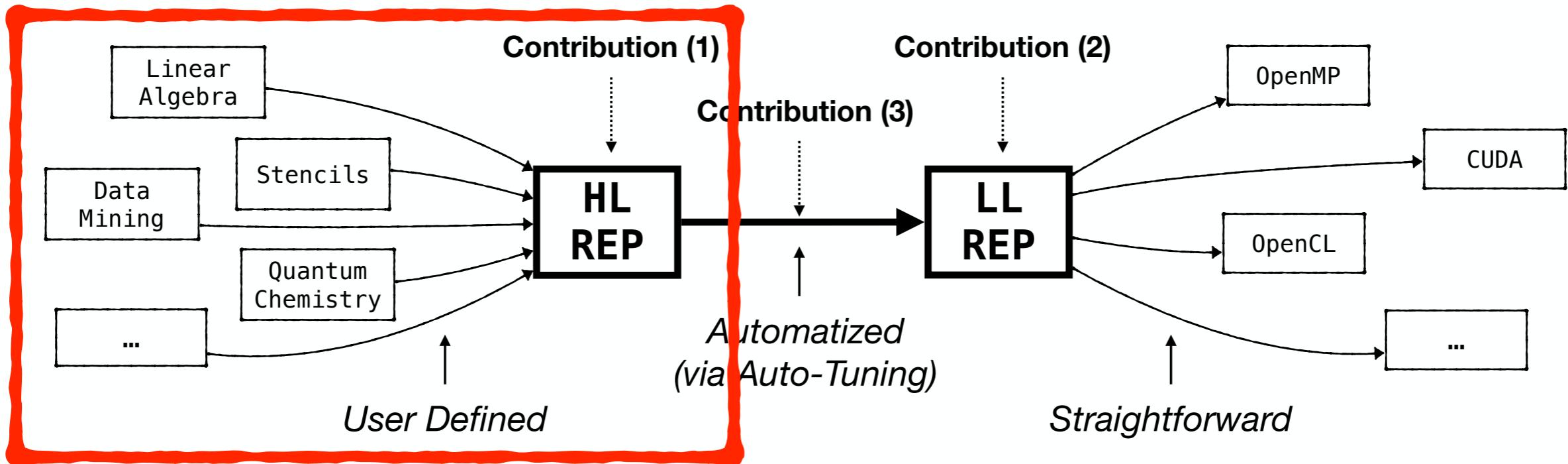
A full version of this article is provided by Rasch [2024], which presents our novel concepts with all of their formal details. In contrast to the full version, this article relies on a simplified formal foundation for better illustration and easier understanding. We often refer the interested reader to Rasch [2024] for formal details that should not be required for understanding the basic ideas and concepts of our approach.

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project PPP-DL (470527619).

Author's Contact Information: Ari Rasch (Corresponding author), University of Muenster, Muenster, Germany; e-mail: a.rasch@uni-muenster.de.

Goal of MDH

MDH is a (formal) framework for expressing & optimizing data-parallel computations:



1. **Contribution 1 (HL-REP):** defines *data parallelism*, based on *common algebraic properties of computations* & introduces *higher-order functions* for expressing these computations, agnostic from hardware and optimization details while still capturing high-level information relevant for generating high-performing code
2. **Contribution 2 (LL-REP):** allows *expressing and reasoning about optimizations* for the memory and core hierarchies of contemporary parallel architectures & generalizes these optimizations to apply to arbitrary combinations of data-parallel computations
3. **Contribution 3 (→):** introduces *automated optimization procedures* for arbitrary combinations of data-parallel computations and parallel architectures — that allows *fully automatic optimization (auto-tuning)*

Focus Today

MDH: High-Level Representation

Goals:

1. Uniform:

should be able to express any kind of data-parallel computation, without relying on domain-specific building blocks, extensions, etc.

2. Minimalistic:

should rely on less building blocks to keep language small and simple

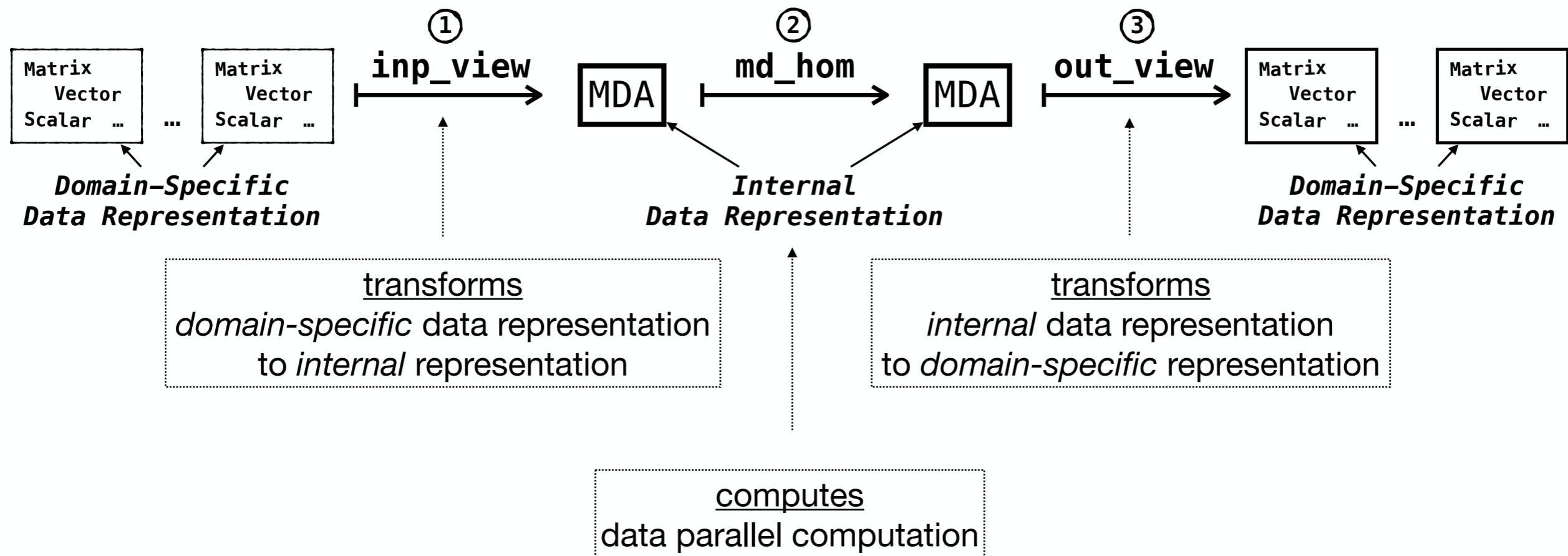
3. Structured:

avoiding compositions and nestings of building blocks as much as possible, thereby further contributing to the usability and simplicity of our language

```
MatVec<T ∈ TYPE | I, K ∈ N> := out_view<T>( w:(i,k) ↦ (i) ) ∘  
                                md_hom<I,K>( *, (#,+) ) ∘  
                                inp_view<T,T>( M:(i,k) ↦ (i,k) , v:(i,k) ↦ (k) )
```

*MDH High-Level Representation of MatVec
(discussed later)*

MDH: High-Level Representation



Our high-level representation (formally) defines data parallelism,
and it expresses these computations using exactly
three straightforwardly composed higher-order functions only

MDH: High-Level Representation

Example: MatVec expressed in MDH

```
MatVec<T∈TYPE| I,K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) ○  
                                md_hom<I,K>( *, (#+,+) ) ○  
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

MDH

MDH High-Level Representation of MatVec

What is happening here:

- `inp_view` captures the accesses to input data
- `md_hom` expresses the data-parallel computation
- `out_view` captures the accesses to output data

```
void MatVec( T[] M, T[] v, T[] w )  
{  
    for( int i=0 ; i < I ; ++i )  
        for( int k=0 ; k < K ; ++k )  
            w[i] += M[i][k] * v[k];  
}
```

MatVec in C++



¹We can generate such MDH expressions also automatically from straightforward (annotated) code in Python, C, ...

MDH: High-Level Representation

	f	\otimes_1	\otimes_2	\otimes_3	\otimes_4
md_hom					
Dot	*	+			
MatVec	*	#	+		
MatMul	*	#	#	+	
MatMul ^T	*	#	#	+	
bMatMul	*	#	#	#	+

Views	inp_view		out_view
	A	B	C
Dot	(k) \mapsto (k)	(k) \mapsto (k)	(k) \mapsto ()
MatVec	(i,k) \mapsto (i,k)	(i,k) \mapsto (k)	(i,k) \mapsto (i)
MatMul	(i,j,k) \mapsto (i,k)	(i,j,k) \mapsto (k,j)	(i,j,k) \mapsto (i,j)
MatMul ^T	(i,j,k) \mapsto (k,i)	(i,j,k) \mapsto (j,k)	(i,j,k) \mapsto (j,i)
bMatMul	(b,i,j,k) \mapsto (b,i,k)	(b,i,j,k) \mapsto (b,k,j)	(b,i,j,k) \mapsto (b,i,j)

1) Linear Algebra Routines

Views	inp_view		out_view
	I	F	O
Conv2D	(p,q,r,s) \mapsto (p+r,q+s)	(p,q,r,s) \mapsto (r,s)	(p,q,r,s) \mapsto (p,q)
MCC	(n,p,...) \mapsto (n,p+r,q+s,c)	(n,p,...) \mapsto (k,r,s,c)	(n,p,...) \mapsto (n,p,q,k)
MCC_Capsule	(n,p,...) \mapsto (n,p+r,q+s,c,mi,mk)	(n,p,...) \mapsto (k,r,s,c,mi,mj)	(n,p,...) \mapsto (n,p,q,k,mi,mj)

2) Convolution Stencils

	f	\otimes_1	...	\otimes_6	\otimes_7
CCSD(T)	*	#	...	#	+

Views	inp_view		out_view
	A	B	C
I1	(a,...,g) \mapsto (g,d,a,b)	(a,...,g) \mapsto (e,f,g,c)	(a,...,g) \mapsto (a,...,f)
I2	(a,...,g) \mapsto (g,d,a,c)	(a,...,g) \mapsto (e,f,g,b)	(a,...,g) \mapsto (a,...,f)

3) Quantum Chemistry

	f	\otimes_1	\otimes_2
PRL	wght	#	max _{PRL}

Views	inp_view		out_view
	N	E	M
PRL	(i,j) \mapsto (i)	(i,j) \mapsto (j)	(i,j) \mapsto (i)

5) Probabilistic Record Linkage

	f	\otimes_1
map(f)	f	#
reduce(\oplus)	id	\oplus
reduce(\oplus, \otimes)	(x) \mapsto (x,x)	(\oplus, \otimes)

Views	inp_view		out_view
	I	0_1	0_2
map(f)	(i) \mapsto (i)	(i) \mapsto (i)	
reduce(\oplus)	(i) \mapsto (i)	(i) \mapsto ()	
reduce(\oplus, \otimes)	(i) \mapsto (i)	(i) \mapsto ()	(i) \mapsto ()

7) Map/Reduce Patterns

	f	\otimes_1	\otimes_2
Histo	f_{Histo}	+	#
GenHisto	f	\oplus	#

Views	inp_view		out_view
	Elems	Bins	Out
Histo	(e,b) \mapsto (e)	(e,b) \mapsto (b)	(e,b) \mapsto (b)
GenHisto	(e,b) \mapsto (e)	(e,b) \mapsto (b)	(e,b) \mapsto (b)

6) Histogram

	f	\otimes_1	\otimes_2
scan(\oplus)	id	# _{prefix-sum} (\oplus)	
MBBS	id	# _{prefix-sum} ($+$)	+

8) Prefix Sum Computations

Views	inp_view		out_view
	A	Out	Out
scan(\oplus)	(i) \mapsto (i)	(i) \mapsto (i)	(i) \mapsto (i)
MBBS	(i,j) \mapsto (i,j)	(i,j) \mapsto (i,j)	(i,j) \mapsto (i,j)

The MDH high-level representation is capable of expressing various kinds of data-parallel computations (with significantly different characteristics)

MDH: High-Level Representation

Differences to existing representations:

In a nutshell

```
MatVec<T ∈ TYPE | I, K ∈ N> := out_view<T>( w:(i,k) ↦ (i) ) ∘
    md_hom<I,K>(*, (+,+)) ∘
    inp_view<T,T>( M:(i,k) ↦ (i,k) , v:(i,k) ↦ (k) )
```

MDH

MatVec in MDH

MDH explicitly captures the operations for combining intermediate results (e.g., add)

MDH separates the scalar operation (e.g., mul) from the operations for combining intermediate results (e.g., add)

The additional semantic information & their explicit separation allows MDH to achieve higher “Performance” & “Expressivity” [1]



```
#map1 = affine_map<(d0, d1) -> (d0, d1)>
#map2 = affine_map<(d0, d1) -> (d1)      >
#map3 = affine_map<(d0, d1) -> (d0)      >
module {
  func.func @main() {
    %M = memref.alloc() : memref<128x64xf32>
    %v = memref.alloc() : memref<64xf32>
    %w = memref.alloc() : memref<128xf32>
    linalg.generic
    {
      indexing_maps = [#map1, #map2, #map3],
      iterator_types = ["parallel", "reduction"]
    } ins(%M,%v:memref<128x64xf32>,memref<64xf32>)
    outs(%w:memref<128xf32>) {
      ^bb0(%in_1: f32, %in_2: f32, %out: f32):
        %0 = arith.mulf %in_1, %in_2 : f32
        %1 = arith.addf %out, %0 : f32
        linalg.yield %1 : f32
    }
    return
  }
}
```

[1] "Linalg vs MDH: A Comparison of two MLIR Dialects", EuroLLVM'24

MatVec in linalg

MDH: High-Level Representation

Differences to existing representations:

```
MatVec<T∈TYPE| I,K∈ℕ> := out_view<T>( w:(i,k)→(i) ) ∘
                                md_hom<I,K>(*, (+,+)) ∘
                                inp_view<T,T>( M:(i,k)→(i,k) , v:(i,k)→(k) )
```

MDH

MatVec in MDH

MDH allows using
different combine operators and has
(formally defined) semantics for them

```
def MatVec (I , K ):
    M = te.placeholder( ( I , K ) , name ='M' , dtype =' float32' )
    v = te.placeholder( ( K ,) , name ='v' , dtype =' float32' )

    k = te.reduce_axis ( (0,K) , name ='k' )
    w = te.compute (
        (I ,),
        lambda i : te.sum( M[i,k] * v[k] , axis=k )
    )
    return [ M , v , w ]
```



MatVec in TVM

MDH allows using **arbitrary combine operators**, whereas TVM
is limited to: min, max, sum, ...

MDH's stronger support for combine operators
contributes to high “*Expressivity*”
(e.g., for examples PRL & MBBS)



tqchen

This is a restriction in the current tensor expression language, because reduction is quite complicated to be processed in nested form.

There are ongoing effort to enhance low-level IR passes to enable more powerful tensorization, which hopefully will resolve the issue you raised.

Feb '20

MDH: High-Level Representation

We offer a **Python interface** for MDH's high-level program representation:

```
MatVec<T∈TYPE | I,K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) o
                                md_hom<I,K>( *, (#+,+) ) o
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

MDH

in MDH Formalism


MatVec


```
def matvec( T: BasicType, I: int, K: int ):

    @mdh()
    def mdh_matvec():
        return (
            out_view[T]( w = [lambda i,k: (i)] ),
            md_hom[I,K]( mul, ( cc, pw(scalar_plus) ) ),
            inp_view[T,T]( M = [lambda i,k: (i,k)] ,
                           v = [lambda i,k: (k) ] )
        )
```



The MDH-Python-Interface is designed
to be very close to MDH's formal representation

MDH: High-Level Representation

We offer a Python interface for MDH's high-level program representation:

```
MatVec<T∈TYPE| I,K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) o MDH
                                md_hom<I,K>( *, ( +, + ) ) o
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

in MDH Formalism

MatVec

```
def matvec( T: BasicType, I: int, K: int ):

    @mdh()
    def mdh_matvec():
        @scalar_function()
        def mul( out: Scalar[T], inp: Scalar[T,T] ):
            out['w'] = inp['M'] * inp['v']

        return (
            out_view[T]( w = [lambda i,k: (i)] ),
            md_hom[I,K]( mul, ( cc, pw(scalar_plus) ) ),
            inp_view[T,T]( M = [lambda i,k: (i,k)] ,
                           v = [lambda i,k: (k) ] )
        )

    return mdh_matvec
```



We allow custom scalar functions

MDH: High-Level Representation

We offer a Python interface for MDH's high-level program representation:

```
MatVec<T∈TYPE| I,K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) o MDH
                                md_hom<I,K>( *, (#+,+) ) o
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

in MDH Formalism

MatVec

```
def matvec( T: BasicType, I: int, K: int ):

    @mdh()
    def mdh_matvec():

        @pw_custom_func()
        def scalar_plus( res: Scalar[T],
                         lhs: Scalar[T], rhs: Scalar[T] ):
            res['w'] = lhs['w'] + rhs['w']

        return (
            out_view[T]( w = [lambda i,k: (i)] ),
            md_hom[I,K]( mul, ( cc, pw(scalar_plus) ) ),
            inp_view[T,T]( M = [lambda i,k: (i,k)] ,
```



We allow custom point-wise operators

MDH: High-Level Representation

We offer a Python interface for MDH's high-level program representation:

```
MatVec<T∈TYPE| I, K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) o
                                md_hom<I,K>(*, (+, +)) o
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

MDH

in MDH Formalism

MatVec

def matvec(T: BasicType, I: int, K: int):

```
@mdh()
def mdh_matvec():
    def cc(T: BasicType, D: int, d: int):
        @combine_operator()
        index_set_function = IndexSetFunction.id,
        scalar_type        = T,
        dimensionality     = D,
        operating_dimension = d
    )
    def cc__T_D_d( I, PQ ):
        def cc__T_D_d__I_PQ( res, lhs,rhs ):
            for i[1, ..., d - 1] in I[1, ..., d - 1]:
                for i[d + 1, ..., D] in I[d + 1, ..., D]:
                    for i[d] in P:
                        res[i[1, ..., d, ..., D]] = lhs[i[1, ..., d, ..., D]]
                    for i[d] in Q:
                        res[i[1, ..., d, ..., D]] = rhs[i[1, ..., d, ..., D]]

        return cc__T_D_d__I_PQ
    return cc__T_D_d

    return (
        out_view[T]( w = [lambda i,k: (i)] ),
        md_hom[I,K]( mul, ( cc, pw(scalar_plus) ) ),
        inp_view[T,T]( M = [lambda i,k: (i,k)] ,
                      v = [lambda i,k: (k) ] )
    )
)
```



We allow custom combine operators

MDH: High-Level Representation

We offer a Python interface for MDH's high-level program representation:

```
MatVec<T∈TYPE| I, K∈ℕ> := out_view<T>( w:(i,k)↦(i) ) o
                                md_hom<I,K>(*, (+, +)) o
                                inp_view<T,T>( M:(i,k)↦(i,k) , v:(i,k)↦(k) )
```

in MDH Formalism

MatVec

```
def matvec(T: BasicType, I: int, K: int):
    @mdh( out( w = Buffer[T] ) ,
          inp( M = Buffer[T], v = Buffer[T] ) ,
          combine_ops( cc, pw(scalar-plus) ) )
def mdh_matvec(w, M, v):
    for i in range(I):
        for k in range(K):
            w[i] = M[i, k] * v[k]
```



**MDH also takes as input
straightforward sequential Python code
(alternative to DSL programs)**



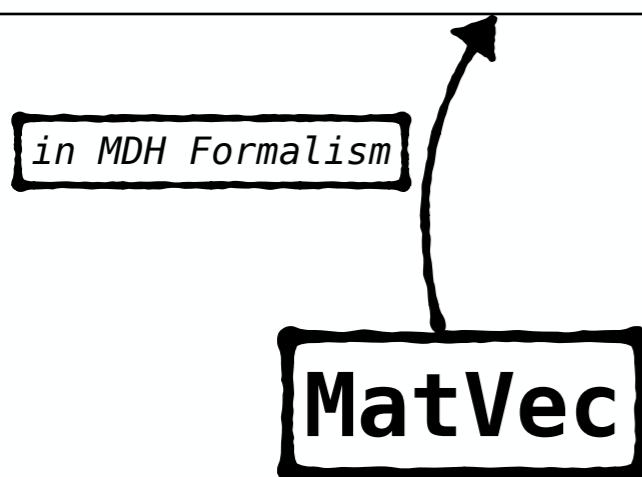
MDH: High-Level Representation

Quick Dive

We offer an **MLIR interface** for MDH's high-level program representation:

```
out_view<f32>( w:(i,k)↔(i) )
  md_hom<128,64>( *, (+,+) )
  inp_view<f32,f32>( M:(i,k)↔(i,k) , v:(i,k)↔(k) )
```

MDH



```
func.func @main()
{
  %M = memref.alloc() : memref<128x64xf32>
  %v = memref.alloc() : memref<64xf32>

  %w = mdh.compute "mdh_matvec"
  {
    inp_view =
    [
      [ affine_map<( i,k ) -> ( i,k )> ],
      [ affine_map<( i,k ) -> ( k ) > ]
    ],
    md_hom =
    {
      scalar_func = @mul,
      combine_ops = [ "cc", ["pw", @add] ]
    },
    out_view =
    [
      [ affine_map<( i,k ) -> ( i )> ]
    ]
  }
  inp_types = [ f32, f32 ],
  mda_size = [ 128, 64 ],
  out_types = [ f32 ]
}(%M,%v):( memref<128x64xf32>,memref<64xf32> )
-> memref<128xf32>

return
```

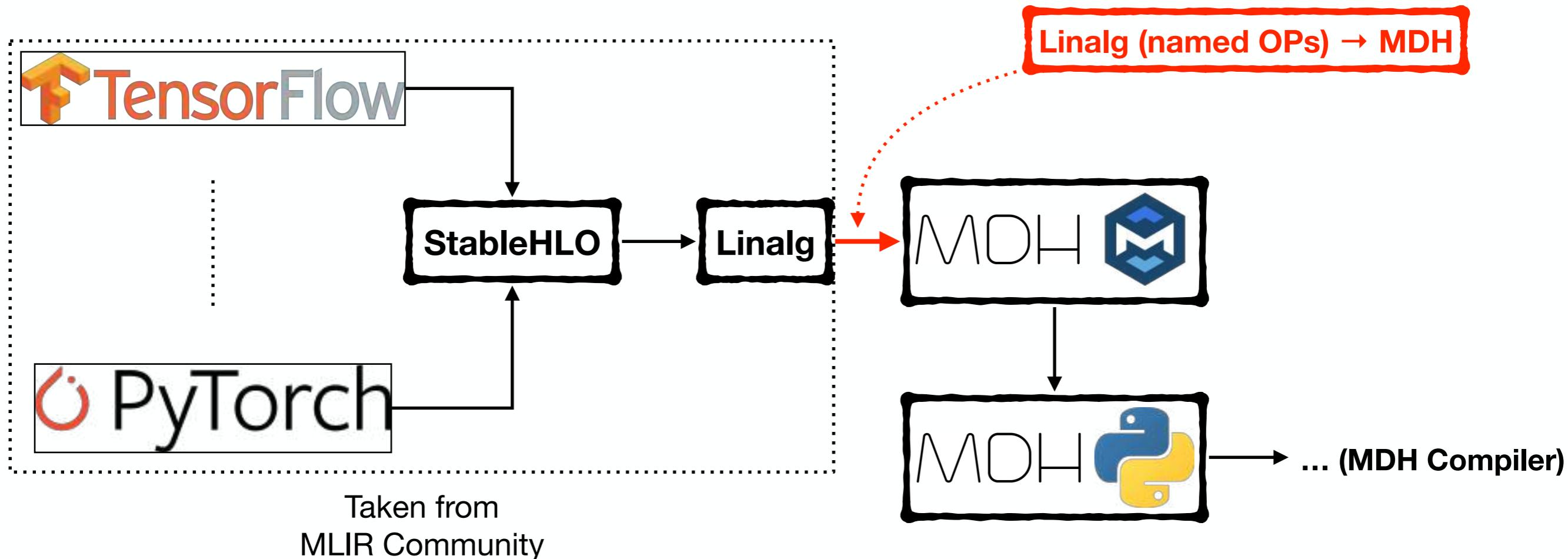


MLIR interface
for MDH



MDH: High-Level Representation

Quick Dive



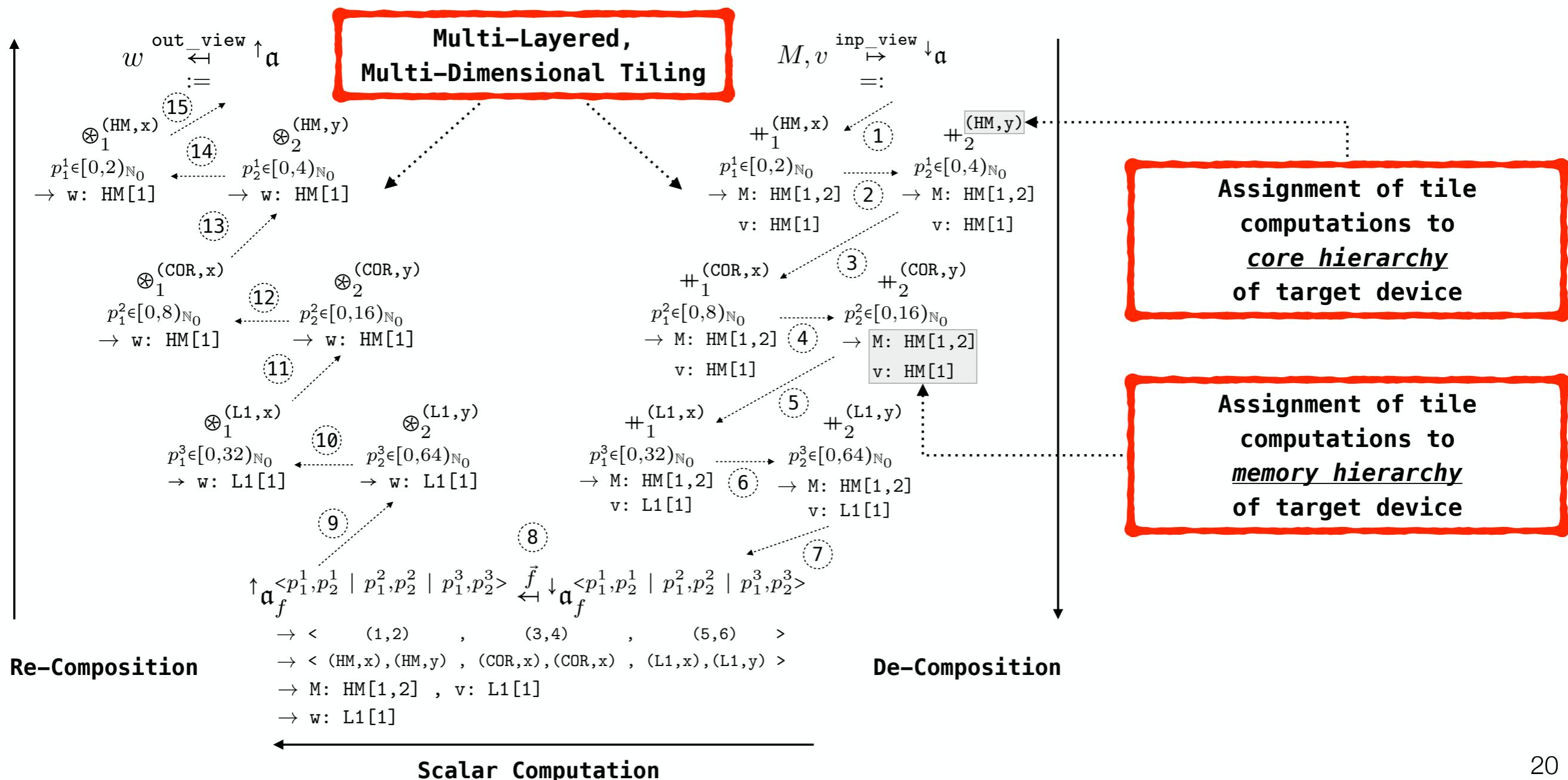
Our MLIR interface allows
easy integration of MDH into DL Frameworks

MDH: Low-Level Representation

Basic Idea

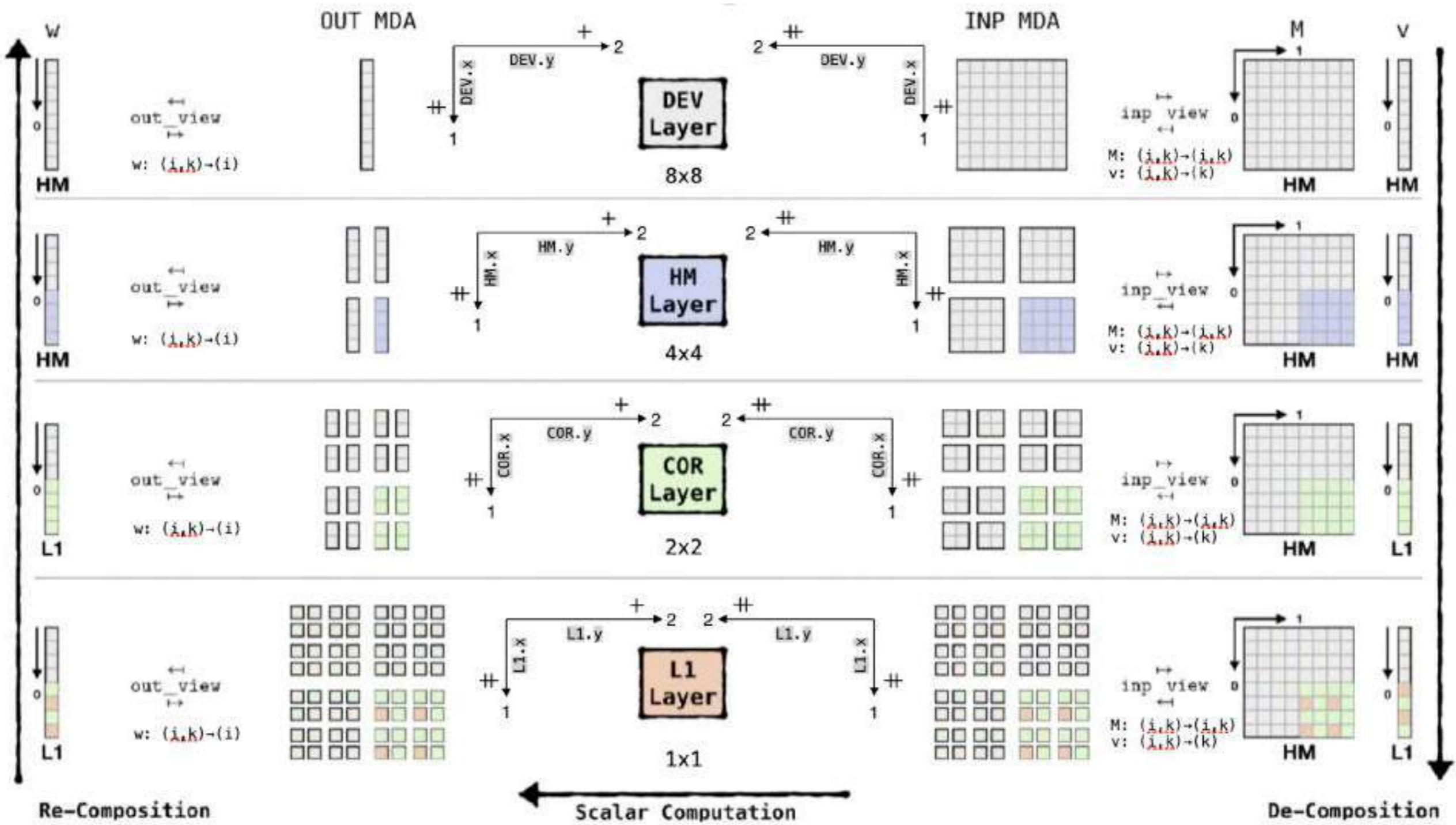
Goals:

1. Expressing a hardware- & data-optimized *de-composition* and *re-composition* of data-parallel computations, based on an *Abstract System Model (ASM)*
 2. Being straightforwardly transformable to executable program code (e.g., in OpenMP, CUDA, and OpenCL) – major optimization decisions explicitly expressed in low-level representation



MDH: Low-Level Representation

Excursion: visualizing low-level instances



The (structured) design of our Low-Level Representation allows uniformly visualizing optimizations

MDH: Lowering: High Level → Low-Level

Based on (formally defined) performance-critical parameters, for a structured optimization process:

No.	Name	Range	Description
0	#PRT	MDH-LVL → \mathbb{N}	number of parts
D1	$\sigma_{\downarrow\text{-ord}}$	MDH-LVL ↔ MDH-LVL	de-composition order
D2	$\leftrightarrow_{\downarrow\text{-ass}}$	MDH-LVL ↔ ASM-LVL	ASM assignment (de-composition)
D3	$\downarrow\text{-mem}^{<\text{ib}>}$	MDH-LVL → MR	memory regions of input BUFs (ib)
D4	$\sigma_{\downarrow\text{-mem}}^{<\text{ib}>}$	MDH-LVL → $[1, \dots, D_{\text{ib}}^{\text{IB}}]_S$	memory layouts of input BUFs (ib)
S1	$\sigma_{f\text{-ord}}$	MDH-LVL ↔ MDH-LVL	scalar function order
S2	$\leftrightarrow_{f\text{-ass}}$	MDH-LVL ↔ ASM-LVL	ASM assignment (scalar function)
S3	$f\downarrow\text{-mem}^{<\text{ib}>}$	MR	memory region of input BUF (ib)
S4	$\sigma_{f\downarrow\text{-mem}}^{<\text{ib}>}$	$[1, \dots, D_{\text{ib}}^{\text{IB}}]_S$	memory layout of input BUF (ib)
S5	$f\uparrow\text{-mem}^{<\text{ob}>}$	MR	memory region of output BUF (ob)
S6	$\sigma_{f\uparrow\text{-mem}}^{<\text{ob}>}$	$[1, \dots, D_{\text{ob}}^{\text{OB}}]_S$	memory layout of output BUF (ob)
R1	$\sigma_{\uparrow\text{-ord}}$	MDH-LVL ↔ MDH-LVL	re-composition order
R2	$\leftrightarrow_{\uparrow\text{-ass}}$	MDH-LVL ↔ ASM-LVL	ASM assignment (re-composition)
R3	$\uparrow\text{-mem}^{<\text{ob}>}$	MDH-LVL → MR	memory regions of output BUFs (ob)
R4	$\sigma_{\uparrow\text{-mem}}^{<\text{ob}>}$	MDH-LVL → $[1, \dots, D_{\text{ob}}^{\text{OB}}]_S$	memory layouts of output BUFs (ob)

*exploiting core hierarchy
(parallelization)*

*exploiting memory hierarchy
(data movements)*

Our parameters
unify & generalize & combine
(and also formalize)
well-proven optimizations
(e.g., tiling, data movements,
and parallelization)

Table 1. Tuning parameters of our low-level expressions

We use our **Auto-Tuning Framework (ATF)** to automatically determine optimized values of parameters¹

¹ We optionally allow (expert) users to incorporate their knowledge into the optimization process via *MDH-Based Schedules* [CC'23]

MDH: Experimental Results

MDH is experimentally evaluated in terms of **Performance & Portability & Productivity**:

Competitors:

1. Scheduling Approach:

- Apache TVM [1] (GPU & CPU)

2. Polyhedral Compilers:

- PPCG [2] (GPU)
- Pluto [3] (CPU)

3. Functional Approach:

- Lift [4] (GPU & CPU)

4. Domain-Specific Libraries:

- NVIDIA cuBLAS & cuDNN (GPU)
- Intel oneMKL & oneDNN (CPU)

Case Studies:

1. Linear Algebra Routines:

- Matrix Multiplication (MatMul)
- Matrix-Vector Multiplication (MatVec)

2. Stencil Computations:

- Jacobi Computation (Jacobi1D)
- Gaussian Convolution (Conv2D)

3. Quantum Chemistry:

- Coupled Cluster (CCSD(T))

4. Data Mining:

- Probabilistic Record Linkage (PRL)

5. Deep Learning:

- Multi-Channel Convolution (MCC)
- Capsule-Style Convolution (MCC_Capsule)

[1] Chen et al., “TVM: An Automated End-to-End Optimizing Compiler for Deep Learning”, OSDI’18

[2] Verdoolaege et al., “Polyhedral Parallel Code Generation for CUDA”, TACO’13

[3] Bondhugula et al., “PLuTo: A Practical and Fully Automatic Polyhedral Program Optimization System”, PLDI’08

[4] Steuwer et al., “Generating Performance Portable Code using Rewrite Rules”, ICFP’15



MDH: Experimental Results

Performance Evaluation: (via runtime comparison)

Highlights only

Deep Learning	NVIDIA Ampere GPU									
	ResNet-50				VGG-16				MobileNet	
	Training		Inference		Training		Inference		Training	Inference
	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MCC
TVM+Ansor	1.00	1.26	1.05	2.22	0.93	1.42	0.88	1.14	0.94	1.00
PPCG	3456.16	8.26	-	7.89	1661.14	7.06	5.77	5.08	2254.67	7.55
PPCG+ATF	3.28	2.58	13.76	5.44	4.26	3.92	9.46	3.73	3.31	10.71
cuDNN	0.92	-	1.85	-	1.22	-	1.94	-	1.81	2.14
cuBLAS	-	1.58	-	2.67	-	0.93	-	1.04	-	-
cuBLASEx	-	1.47	-	2.56	-	0.92	-	1.02	-	-
cuBLASLtt	-	1.26	-	1.22	-	0.91	-	1.01	-	-



NVIDIA

NVCC vs NVRTC

MDH speedup over

- TVM: 0.88x – 2.22x
- PPCG: 2.58x – 13.76x
- (cuBLAS/cuDNN: 0.91x – 2.67x)

Deep Learning	Intel Skylake CPU									
	ResNet-50				VGG-16				MobileNet	
	Training		Inference		Training		Inference		Training	Inference
	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MCC
TVM+Ansor	1.53	1.05	1.14	1.20	1.97	1.14	2.38	1.27	3.01	1.40
Pluto	355.81	49.57	364.43	13.93	130.80	93.21	186.25	36.30	152.14	75.37
Pluto+ATF	13.08	19.70	170.69	6.57	3.11	6.29	53.61	8.29	3.50	25.41
oneDNN	0.39	-	5.07	-	1.22	-	9.01	-	1.05	4.20
oneMKL	-	0.44	-	1.09	-	0.88	-	0.53	-	-
oneMKL(JIT)	-	6.43	-	8.33	-	27.09	-	9.78	-	-



MDH speedup over

- TVM: 1.05 – 3.01x
- Pluto: 6.29x – 364.43x
- (oneMKL/oneDNN: 0.39x – 9.01x)

Case Study “Deep Learning” for which most competitors are highly optimized (most challenging for us!)

Significantly higher speedups for other case studies,
e.g., >170x over TVM on GPU already for straightforward dot products

MDH: Experimental Results

Portability Evaluation: (via Pennycook Metric [6])

Highlights only

Deep Learning	Pennycook Metric									
	ResNet-50				VGG-16				MobileNet	
	Training		Inference		Training		Inference		Training	Inference
	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MatMul	MCC	MCC
MDH+ATF	0.67	0.76	0.91	1.00	0.98	0.95	0.97	0.68	0.98	1.00
TVM+Ansor	0.53	0.62	0.89	0.59	0.76	0.81	0.70	0.61	0.54	0.75

The other related approaches achieve lowest portability — of “0.00” — only, because they are designed for particular architectures and/or application classes only

MDH: Experimental Results

Highlights only

Productivity Evaluation: (via intuitive argumentation)

```
1 cublasSgemv( /* ... */ );
```

Listing 4. cuBLAS program expressing Matrix-Vector Multiplication (MatVec)

```
1 for( int i = 0 ; i < M ; ++i )
2     for( int k = 0 ; k < K ; ++k )
3         w[i] += M[i][k] * v[k];
```

Listing 2. PPCG/Pluto program expressing Matrix-Vector Multiplication (MatVec)¹

```
1 def MatVec(I, K):
2     M = te.placeholder((I, K), name='M', dtype='float32')
3     v = te.placeholder((K,), name='v', dtype='float32')
4
5     k = te.reduce_axis((0, K), name='k')
6     w = te.compute(
7         (I,),
8         lambda i: te.sum(M[i, k] * v[k], axis=k)
9     )
10    return [M, v, w]
```

Listing 1. TVM program expressing Matrix-Vector Multiplication (MatVec)

```
1 nFun(n => nFun(m =>
2     fun(matrix: [[float]]n)m => fun(xs: [float]n =>
3         matrix :>> map(fun(row =>
4             zip(xs, row) :>> map(*) :>> reduce(+, 0)
5         )) ) ) )
```

Listing 3. Lift program expressing Matrix-Vector Multiplication (MatVec)

¹MDH can also take (annotated) sequential code as input [IMPACT'20]

MDH: WIP & Future Directions

Many promising future directions (detailed discussion available [here](#)):

High-Level Program Transformations (a.k.a. Fusion)

`md_hom(g,(+,...,+)) o md_hom(f,(+,...,+))`

→ `md_hom(g o f, (+,...,+))`

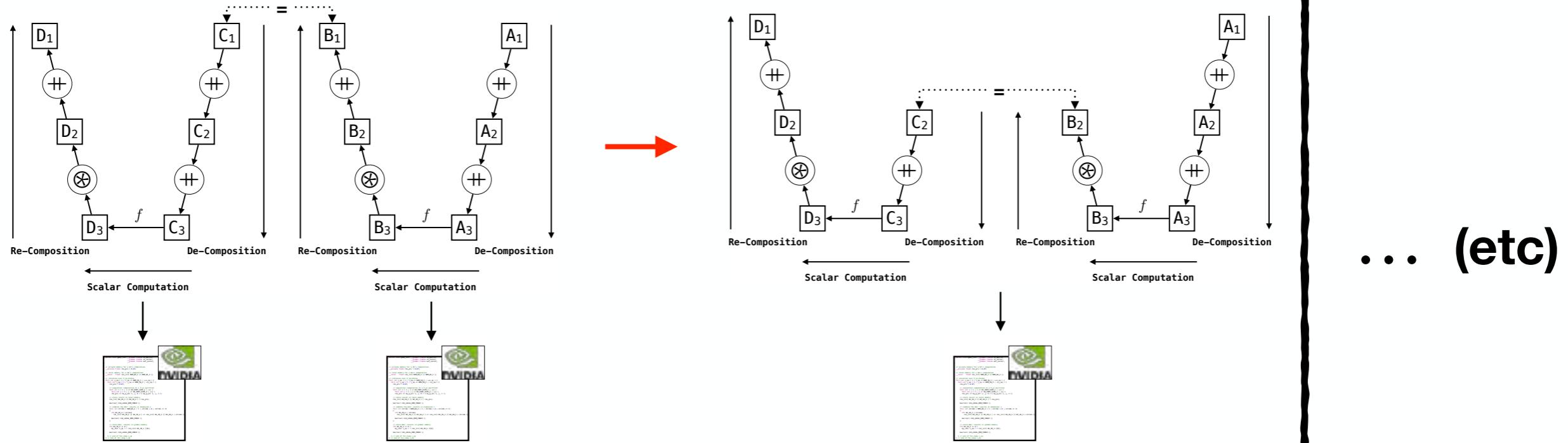
Sparse Computations

...

Domain-Specific HW

...

Low-Level Program Transformations (a.k.a. Fusion)



We consider MDH to be a promising (formal) foundation for these goals,
e.g., due to its uniform representation and its captured algebraic information

MDH: Summary

- MDH combines three key goals – ***Performance & Portability & Productivity*** – as compared to related approaches
- For this, MDH **formally introduces program representations** on both:
 - **high level**, for conveniently expressing – in one uniform formalism – the various kinds of data-parallel computations, agnostic from hardware and optimization details, while still capturing all information relevant for generating high-performance program code
 - **low level**, which allows uniformly reasoning – in the same formalism – about optimized (de/re)-compositions of data-parallel computations for the memory and core hierarchies of contemporary parallel architectures (GPUs, CPUs, etc)
- MDH **lowers** instances in its high-level representation to device- and data-optimized instances in its low-level representation, in a **formally sound** manner, by introducing a generic search space that is based on **performance-critical parameters & auto-tuning**
- Our **experiments** confirm that MDH often achieves higher ***Performance & Portability & Productivity*** than popular state-of-practice approaches, including hand-optimized libraries provided by vendors
- **Many promising future directions!**

MDH: Summary

That was a very quick, informal dive!

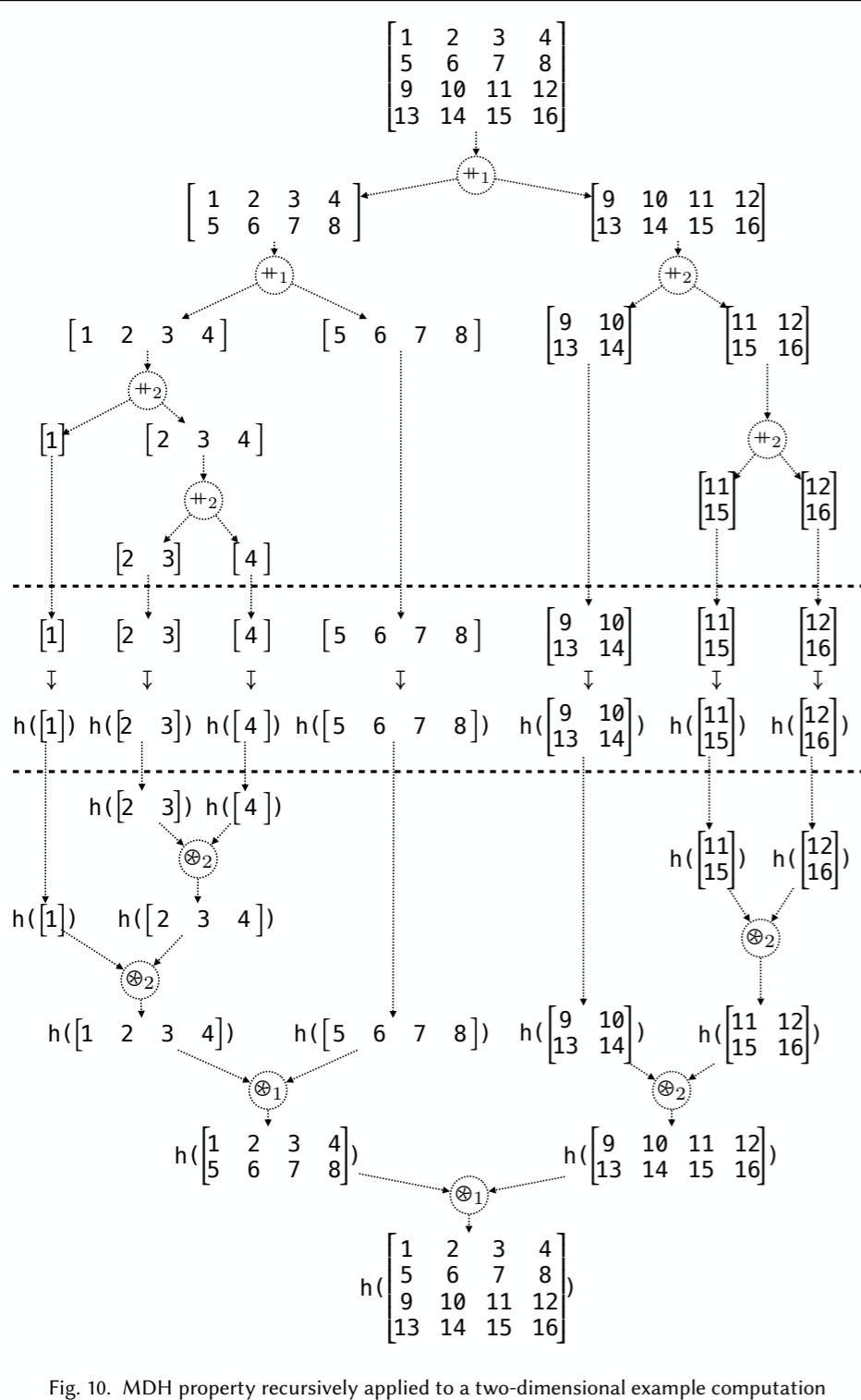


Fig. 10. MDH property recursively applied to a two-dimensional example computation

Definition 3 (Multi-Dimensional Homomorphism). Let $T^{\text{INP}}, T^{\text{OUT}} \in \text{TYPE}$ be two arbitrary scalar types, $D \in \mathbb{N}$ a natural number, and $\Rightarrow_{\text{MDA}}^1, \dots, \Rightarrow_{\text{MDA}}^D : \text{MDA-IDX-SETS} \rightarrow \text{MDA-IDX-SETS}$ functions on MDA index sets. Let further $++_d := ++^{< T^{\text{INP}} | D | d } \in \text{CO}^{< id | T^{\text{INP}} | D | d }$ denote concatenation (Definition 1) in dimension $d \in [1, D]_{\mathbb{N}}$ on D -dimensional MDAs that have scalar type T^{INP} .

A function

$$h^{< I_1, \dots, I_D \in \text{MDA-IDX-SETS} >} : T^{\text{INP}}[I_1, \dots, I_D] \rightarrow T^{\text{OUT}}[\Rightarrow_{\text{MDA}}^1(I_1), \dots, \Rightarrow_{\text{MDA}}^D(I_D)]$$

is a *Multi-Dimensional Homomorphism (MDH)* that has *input scalar type* T^{INP} , *output scalar type* T^{OUT} , *dimensionality* D , and *index set functions* $\Rightarrow_{\text{MDA}}^1, \dots, \Rightarrow_{\text{MDA}}^D$, iff for each $d \in [1, D]_{\mathbb{N}}$, there exists a combine operator $\otimes_d \in \text{CO}^{< \Rightarrow_{\text{MDA}}^D | T^{\text{OUT}} | D | d }$ (Definition 2), such that for any concatenated input MDA $a_1 ++_d a_2$ in dimension d , the *homomorphic property* is satisfied:

$$h(a_1 ++_d a_2) = h(a_1) \otimes_d h(a_2)$$

We denote the type of MDHs concisely as $\text{MDH}^{< T^{\text{INP}}, T^{\text{OUT}} | D | (\Rightarrow_{\text{MDA}}^d)_{d \in [1, D]_{\mathbb{N}}} >}$.

Definition 5 (Buffer). Let $T \in \text{TYPE}$ be an arbitrary scalar type, $D \in \mathbb{N}_0$ a natural number⁹, and $N := (N_1, \dots, N_D) \in \mathbb{N}^D$ a sequence of natural numbers.

A *Buffer (BUF)* b that has *dimensionality* D , *size* N , and *scalar type* T is a function with the following signature:

$$b : [0, N_1]_{\mathbb{N}_0} \times \dots \times [0, N_D]_{\mathbb{N}_0} \rightarrow T \cup \{\perp\}$$

Here, \perp denotes the *undefined* value. We refer to $[0, N_1]_{\mathbb{N}_0} \times \dots \times [0, N_D]_{\mathbb{N}_0} \rightarrow T \cup \{\perp\}$ as the *type* of BUF b , which we also denote as $T^{N_1 \times \dots \times N_D}$, and we refer to set $\text{BUF-IDX-SETS} := \{[0, N]_{\mathbb{N}_0} \mid N \in \mathbb{N}\}$ as *BUF index sets*. Analogously to Notation 1, we write $b[i_1, \dots, i_D]$ instead of $b(i_1, \dots, i_D)$ to avoid a too heavy usage of parentheses.

Definition 2 (Combine Operator). Let $\text{MDA-IDX-SETS} \times \text{MDA-IDX-SETS} := \{(P, Q) \in \text{MDA-IDX-SETS} \times \text{MDA-IDX-SETS} \mid P \cap Q = \emptyset\}$ denote the set of all pairs of MDA index sets that are disjoint. Let further $\Rightarrow_{\text{MDA}}^D : \text{MDA-IDX-SETS} \rightarrow \text{MDA-IDX-SETS}$ be a function on MDA index sets, $T \in \text{TYPE}$ a scalar type, $D \in \mathbb{N}$ an MDA dimensionality, and $d \in [1, D]_{\mathbb{N}}$ an MDA dimension.

We refer to any binary function \otimes of type (parameters in angle brackets are type parameters)

$$\otimes^{< (I_1, \dots, I_{d-1}, I_d, I_{d+1}, \dots, I_D) \in \text{MDA-IDX-SETS}^{D-1}, (P, Q) \in \text{MDA-IDX-SETS} \times \text{MDA-IDX-SETS} >} :$$

$$T[I_1, \dots, \underbrace{\Rightarrow_{\text{MDA}}^D(P), \dots, I_D}_d] \times T[I_1, \dots, \underbrace{\Rightarrow_{\text{MDA}}^D(Q), \dots, I_D}_d] \rightarrow T[I_1, \dots, \underbrace{\Rightarrow_{\text{MDA}}^D(P \cup Q), \dots, I_D}_d]$$

as *combine operator* that has *index set function* $\Rightarrow_{\text{MDA}}^D$, *scalar type* T , *dimensionality* D , and *operating dimension* d . We denote combine operator's type concisely as $\text{CO}^{< \Rightarrow_{\text{MDA}}^D | T | D | d >}$.

... (>130 pages)

All concepts are fully formally defined in the MDH paper
(arXiv version)



Code Optimization via ATF



Overview Getting Started Code Examples Publications Citations Contact



Auto-Tuning Framework (ATF)

Efficient Auto-Tuning of Parallel Programs with Constrained Tuning Parameters

Overview

The **Auto-Tuning Framework (ATF)** is a general-purpose auto-tuning approach: given a program that is implemented as generic in performance-critical program parameters (a.k.a. *tuning parameters*), such as sizes of tiles and numbers of threads, ATF fully automatically determines a hardware- and data-optimized configuration of such parameters.

Key Feature of ATF

A key feature of ATF is its support for **Tuning Parameter Constraints**. Parameter constraints allow auto-tuning programs whose tuning parameters have so-called *interdependencies* among them, e.g., the value of one tuning parameter has to evenly divide the value of another tuning parameter.

ATF's support for parameter constraints is important: modern parallel programs target novel parallel architectures, and such architectures typically have deep memory and core hierarchies thus requiring constraints on tuning parameters, e.g., the value of a tile size tuning parameter on an upper memory layer has to be a multiple of a tile size value on a lower memory layer.

For such parameters, ATF introduces novel concepts for **Generating & Storing & Exploring** the search spaces of constrained tuning parameters, thereby contributing to a substantially more efficient overall auto-tuning process for such parameters, as confirmed in our **Experiments**.

Generality of ATF

For wide applicability, ATF is designed as generic in:

1. The target program's **Programming Language**, e.g., C/C++, CUDA, OpenMP, or OpenCL. ATF offers *pre-implemented cost functions* for conveniently auto-tuning C/C++ programs, as well as CUDA and OpenCL kernels which require host code for their execution which is automatically generated and executed by ATF's pre-implemented CUDA and OpenCL cost functions. ATF also offers a pre-implemented generic cost function that can be used for conveniently auto-tuning programs in any other programming language different from C/C++, CUDA, and OpenCL.

<https://atf-tuner.org>

ACM TACO 2021

Efficient Auto-Tuning of Parallel Programs with Interdependent Tuning Parameters via Auto-Tuning Framework (ATF)

ARI RASCH and RICHARD SCHULZE, University of Muenster, Germany

MICHEL STEUWER, University of Edinburgh, United Kingdom

SERGEI GORLATCH, University of Muenster, Germany

Auto-tuning is a popular approach to program optimization: it automatically finds good configurations of a program's so-called tuning parameters whose values are crucial for achieving high performance for a particular parallel architecture and characteristics of input/output data. We present three new contributions of the Auto-Tuning Framework (ATF), which enable a key advantage in *general-purpose auto-tuning*: efficiently optimizing programs whose tuning parameters have *interdependencies* among them. We make the following contributions to the three main phases of general-purpose auto-tuning: (1) ATF *generates* the search space of interdependent tuning parameters with high performance by efficiently exploiting parameter constraints; (2) ATF *stores* such search spaces efficiently in memory, based on a novel chain-of-trees search space structure; (3) ATF *explores* these search spaces faster, by employing a multi-dimensional search strategy on its chain-of-trees search space representation. Our experiments demonstrate that, compared to the state-of-the-art, general-purpose auto-tuning frameworks, ATF substantially improves generating, storing, and exploring the search space of interdependent tuning parameters, thereby enabling an efficient overall auto-tuning process for important applications from popular domains, including stencil computations, linear algebra routines, quantum chemistry computations, and data mining algorithms.

CCS Concepts: • General and reference → Performance; • Computer systems organization → Parallel architectures; • Software and its engineering → Parallel programming languages;

Additional Key Words and Phrases: Auto-tuning, parallel programs, interdependent tuning parameters

ACM Reference format:

Ari Rasch, Richard Schulze, Michel Steuwer, and Sergei Gorlatch. 2021. Efficient Auto-Tuning of Parallel Programs with Interdependent Tuning Parameters via Auto-Tuning Framework (ATF). *ACM Trans. Archit. Code Optim.* 18, 1, Article 1 (January 2021), 26 pages.

<https://doi.org/10.1145/3427093>

This is a new paper, not an extension of a conference paper.

Authors' addresses: A. Rasch, R. Schulze, and S. Gorlatch, University of Muenster, Muenster, Germany; emails: {a.rasch, r.schulze, gorlatch}@uni-muenster.de; M. Steuwer, University of Edinburgh, Edinburgh, United Kingdom; email: michel.steuwer@glasgow.ac.uk.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

1544-3566/2021/01-ART1

Goal of ATF

Auto-Tuning Framework (ATF) a major advantage over related approaches:

ATF finds values of performance-critical parameters with
interdependencies among them
via optimized processes to
generating & storing & exploring
the spaces of interdependent parameters

ATF auto-tunes programs written in **arbitrary programming languages**, particularly CUDA.

For this, ATF introduces:

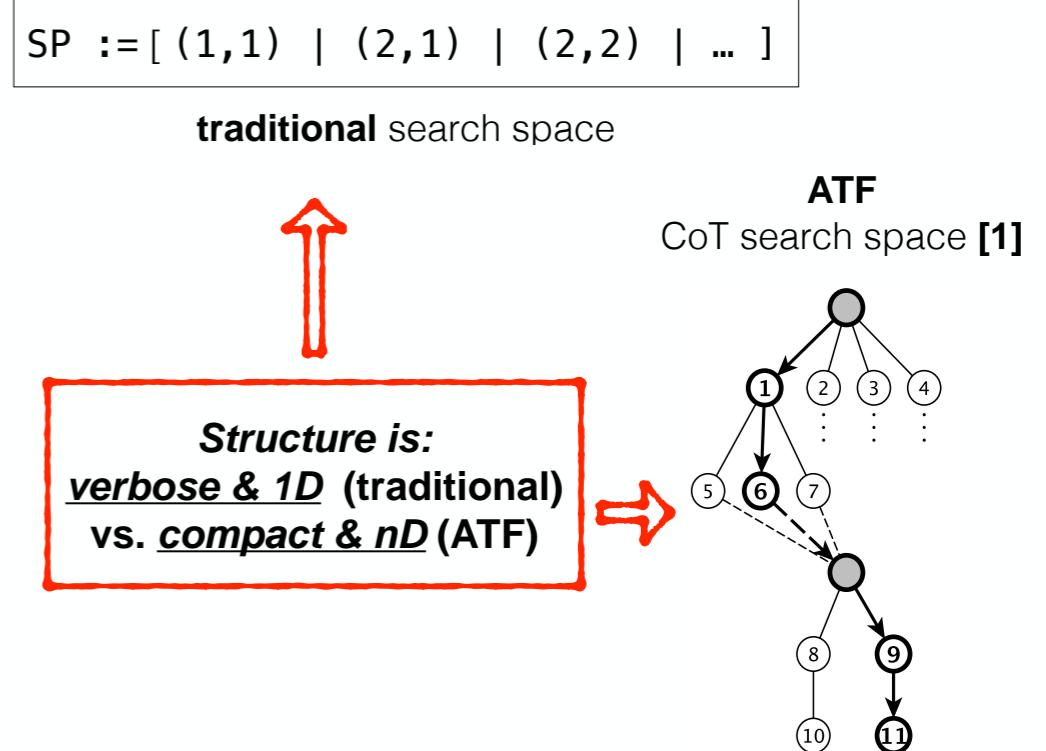
```
tuner.addParameter( "tp_1", T1 );
tuner.addParameter( "tp_2", T2 );
// ...
tuner.addConstraint(
    [](T1 tp_1, T2 tp_2, ... ) -> bool
```

traditional constraints

```
tuner.addParameter( "tp_1", R1, [](T1 tp_1) -> bool { /* ... */ } );
tuner.addParameter( "tp_2", R2, [](T2 tp_2) -> bool { /* ... */ } );
```

ATF
parameter constraints [1]

Defined on:
search space (traditional)
vs. parameters (ATF)



ATF: User Interfaces

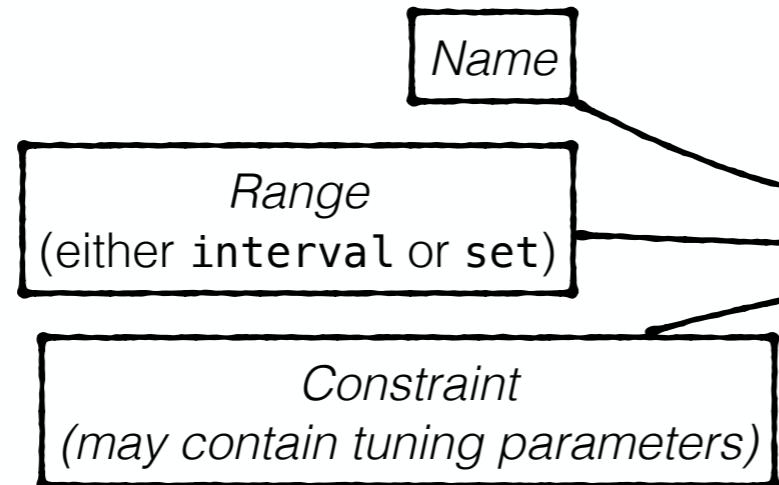
ATF's ***Python-based*** user interface¹:



ATF Website



Schulze, Gorlatch, Rasch, “pyATF: Constraint-Based Auto-Tuning in Python”, CC’25




```

# Input Size
N = 1000

# Step 1: Generate the Search Space
WPT = TP('WPT',
          Interval(1,N),
          lambda WPT: N % WPT == 0)

LS = TP('LS',
         Interval(1,N),
         lambda WPT,LS: (N/WPT) % LS == 0)

# Step 2: Implement a Cost Function
saxpy_kernel = # ... (kernel's code & name)

N = np.int32(N)
a = np.float32(np.random.random())
x = np.random.rand(N).astype(np.float32)
y = np.random.rand(N).astype(np.float32)

cf = cuda.CostFunction(saxpy_cuda_kernel) \
    .device_id(0) \
    .kernel_args(N, a, x, y) \
    .grid_dim(lambda WPT,LS: N/WPT/LS) \
    .block_dim(lambda LS: LS)

# Step 3: Explore the Search Space
config = Tuner().tuning_parameters(WPT,LS) \
    .search_technique(AUCBandit()) \
    .tune(cf, Evaluations(50))
  
```

¹ ATF also offers a GPL-based interface for (online-tuning) C++ programs, as well as a DSL-based interface (offline tuning)

ATF: Summary

ATF efficiently handles tuning parameters with *interdependencies* among them:

ATF introduces novel concepts to
Generating & Storing & Exploring
the search spaces of interdependent parameters, based on its *novel*
constraint design and ***search space representation***

Further ATF features (not presented on slides for brevity):

- ATF has a DSL-based **user interface** that is **arguably simpler** to use and more expressive than existing auto-tuners (including: OpenTuner & CLTune)
- ATF offers different kinds of **search techniques** and **abort conditions** (extensible)
- ATF offers a **DSL-based** user interface (*offline tuning*), as well as **GPL-based** interfaces (*online* auto-tuning):



pyATF

github.com/atf-tuner/pyATF



cppATF

github.com/atf-tuner/cppATF

... (future work)



Universität
Münster

Questions?



<https://arirasch.net>
a.rasch@uni-muenster.de



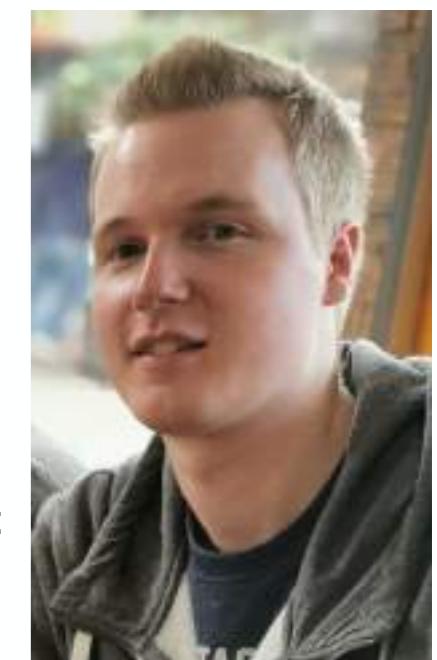
Ari
Rasch



<https://mdh-lang.org>



<https://richardschulze.net>
r.schulze@uni-muenster.de



Code
Generation



<https://atf-tuner.org>

Code
Optimization