Performance, Portability, and Productivity for Data-Parallel Computations on Multi- and Many-Core Architectures

a.rasch@wwu.de

Ari Rasch, Richard Schulze, and Sergei Gorlatch

WWU MÜNSTER

Generation

Let T and T' be two arbitrary types. A function $h : T[N_1]\dots[N_d] \to T'$ on d-dimensional arrays is called a *Multi-Dimensional Homomorphism (MDH)* iff there exist *combine operators* $\circledast_1, \dots, \circledast_d : T' \times T' \to T'$, such that for each $k \in [1, d]$ and arbitrary, concatenated input MDA $a +\!\!+_k b$:

$$h(\ a +\!\!+_k b\ )\ =\ h(a)\ \circledast_k\ h(b)$$

MDHs can be represented uniformly via our md_hom parallel pattern:

$$\texttt{md\_hom}(\ f\ ,\ (\circledast_1, \dots, \circledast_d)\ )(\ a[N_1]\dots[N_d]\ )\ =\ \underset{i_1\in[1,N_1]}{\circledast_1}\ \dots\ \underset{i_d\in[1,N_d]}{\circledast_d}\ f(\ a[i_1]\dots[i_d]\ )$$

Important computations are MDHs:

Linear Algebra (BLAS)

```
GEMM = md_hom( *, (++, ++, +) ) o view(A,B)(i,j,k)(A[i,k],B[k,j])
GEMV = md_hom( *, (++,     +) ) o view(A,B)(i,  k)(A[i,k],B[k]  )
DOT  = md_hom( *, (        +) ) o view(A,B)(    k)(A[k]  ,B[k]  )
```

Data Mining

```
PRL = md_hom( weight, (++, max) ) o view(…)
```

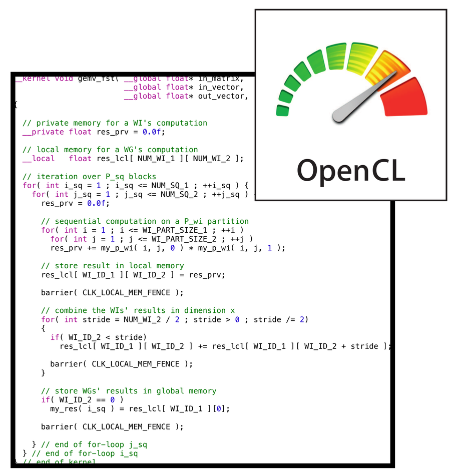
Machine Learning

```
TC = md_hom( *, (++,…,++ , +,…,+) ) o view(…)
```

Stencil Computations

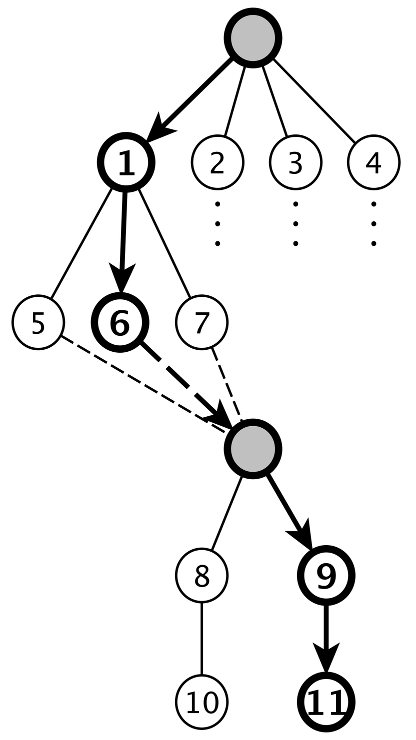
```
Gaussian_2D = md_hom( G_func, (++,++)     ) o view(…)
Jacobi_3D   = md_hom( J_func, (++,++, ++) ) o view(…)
```

$\texttt{md\_hom}(\ f,\ (\circledast_1, \dots, \circledast_k)\ )$ → ***Generating OpenCL Code*** *(auto-tunable)*



Optimization

Our **Auto-Tuning Framework (ATF)** is a **general-purpose approach** that supports auto-tuning of programs written in **arbitrary programming languages** and that may have **interdependent tuning parameters.**



We provide a novel ***chain-of-trees*** *search space structure* for interdependent tuning parameters.

```
#atf::tp name       /* name       */
         range      /* range      */
         constraint /* constraint */
```

We extend the traditional definition of *tuning parameters* by a ***parameter's constraint***.

ATF efficiently *generates*, *stores*, and *explores* the spaces of interdependent tuning parameters

2.75x faster than ***TVM***

1.37x faster than newest ***Intel MKL/NVIDIA cuBLAS***

Our MDH approach shows often **significantly better performance** as compared to the currently best-performing **performance-portable** and **hand-optimized approaches.**

39x faster than ***EKR***

2x faster than ***COGENT & Tensor Comprehensions***