

Georgi Gaydadjiev, Director of Maxeler IoT-Labs BV, Delft Honorary Visiting Professor at the Department of Computing, Imperial College London



Yale: 80 in 2019, Universitat Politècnica de Catalunya, Barcelona, Spain, 1 July 2019

Καλλίπολις – Callipolis also known as Barcelona now



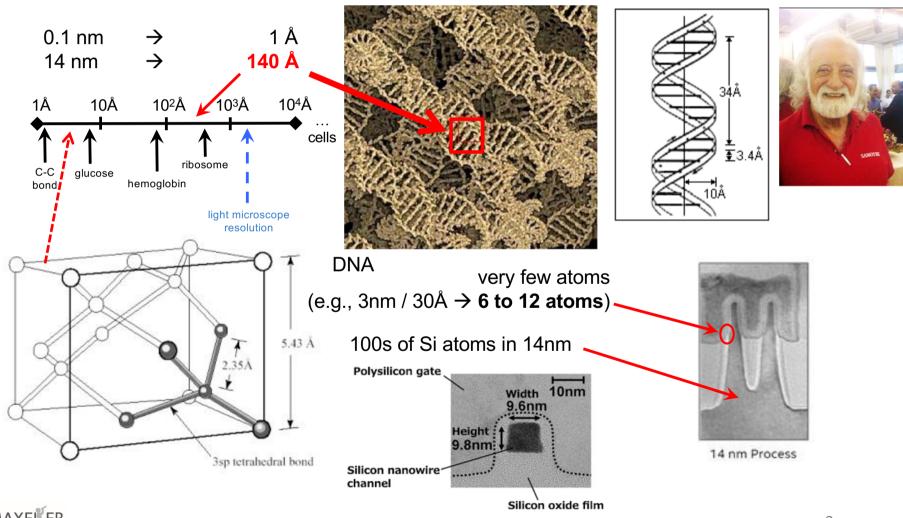
Disclaimer

- Some of the views expressed here are those of the speaker and not his employer
- You probably are already familiar with many of the concepts presented
- Most of the work here is done by others, however, if there is an error it is entirely my fault
- The roadmap of Maxeler may not be in full agreement with some of the statements expressed in this talk
- No copyright infringement is intended

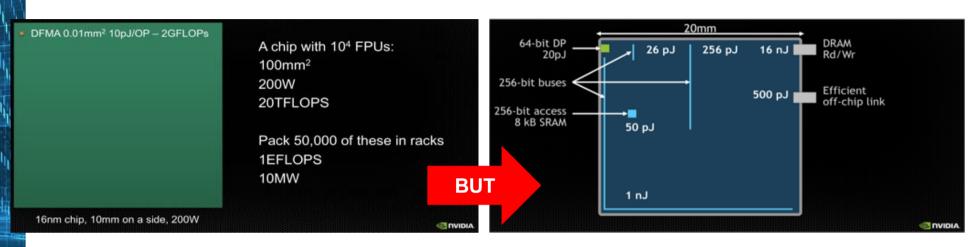


Think Ångströms not nanometers

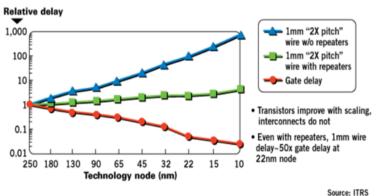
We should steer the movements of almost each individual electron to solve our specific problem



The Data Movement Challenge



Processor Technology	40 nm	10nm
Vdd (nominal)	0.9 V	0.7 V
DFMA energy	50 pJ	7.6 pJ
64b 8 KB SRAM Rd	14 pJ	2.1 pJ
Wire energy (256 bits, 10mm)	310 pJ	174 pJ



(Courtesy: NVIDIA and ITRS)

Moving data off-chip will cost ~200x more energy and is also much slower

Wires that carry the data (and instructions, if any) at all levels should be considered seriously





All started with an Application Specific Computer

1939

- ABC to solve 30 equations simultaneously
- We moved to programmable computers
 - ENIAC and many, many more
- Laws of Physics enforce back Application Specific
 - Custom Accelerators keep emerging
- Laws of Economics demand Programmability
 - Fully Programmable Application Specific Systems?



2019

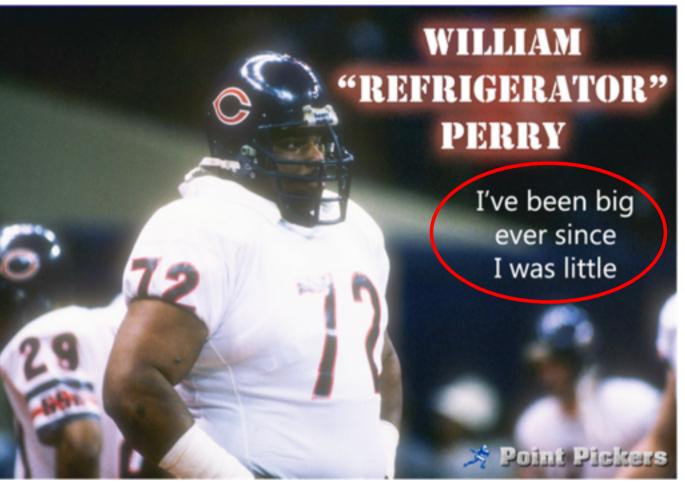
Maybe now we can finally make the *computing machines proper**?

*John V Atanasoff's name for digital computers

It is our job since Quantum Computers schedule shifted with 10-15 years (again)



We know about custom accelerators





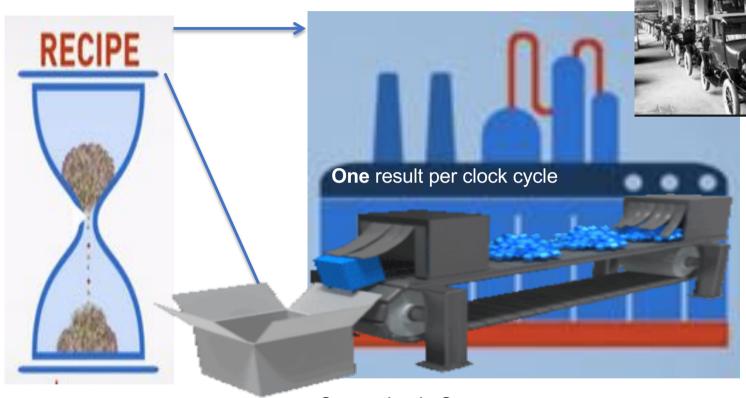
- Dedicated Arithmetic
- Vector Operations
- Data Streaming
- I/O Processing
- Complex Operations (FFT, DCT, ...)
- Complex" Algorithms (Neural Networks, Compression, ...)

Power on the "refrigerator" when a lot of specialist work has to be done.

Can we program the accelerator to be always as big as required for any arbitrary job?



Build Computers for your Problem and Data



Computing in Time:
Follow a recipe step by step
one at the time

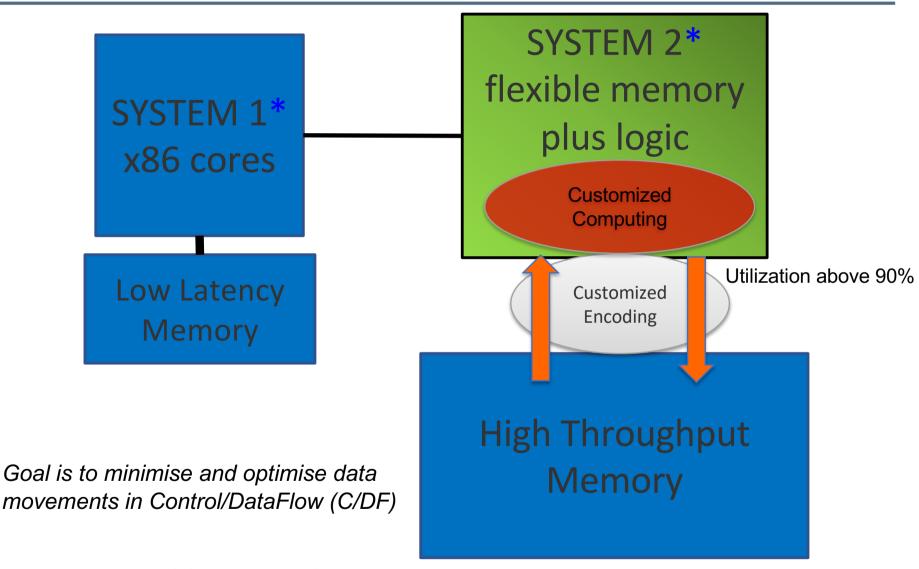
Computing in Space:
Build a "recipe specific" factory with multiple
paths performed simultaneously

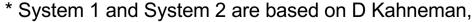
Efficient, predictable, reliable "mass production" of huge data amounts

At each clock tick all data in processing move one stage ahead -> massive throughput



The Combined Control/DataFlow System





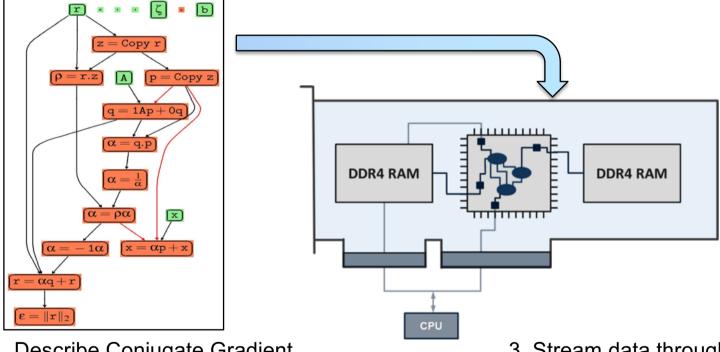
[&]quot;Thinking Fast Thinking Slow", Nobel Prize in Economics, 2002



Programming a Dataflow "mass production" Engine

Create customized mega accelerators with massive inherent throughputs

2. Compile dataflow structure and load to hardware

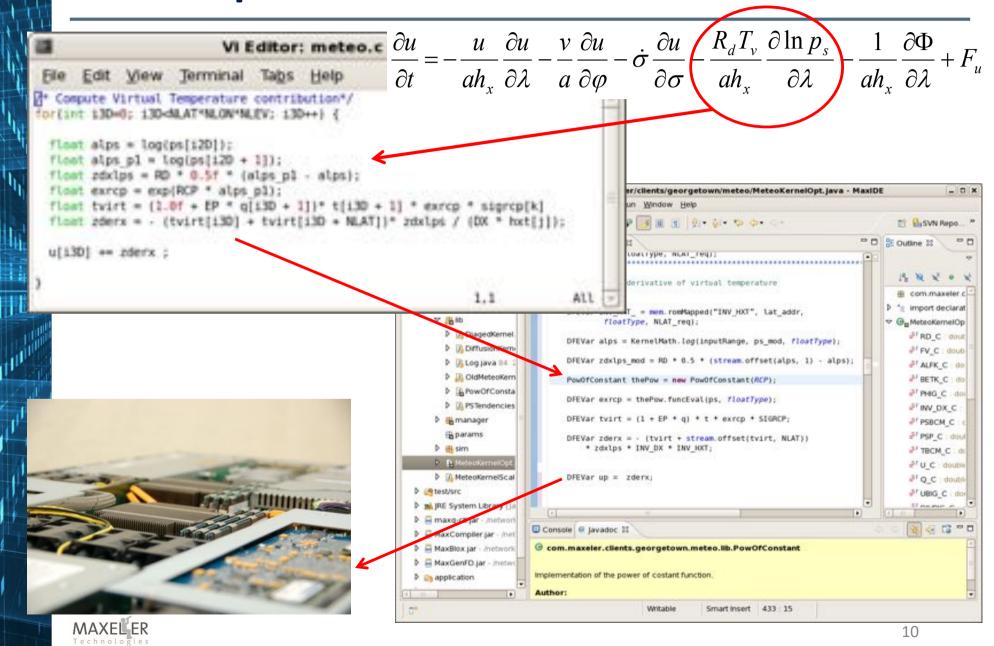


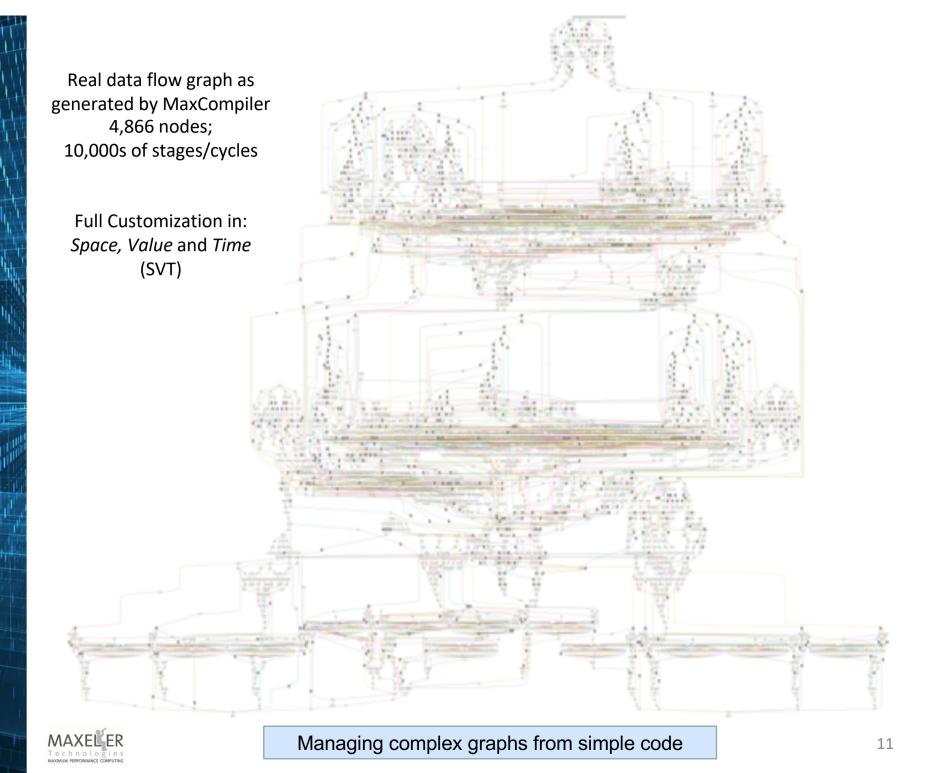


3. Stream data through the Custom Accelerator

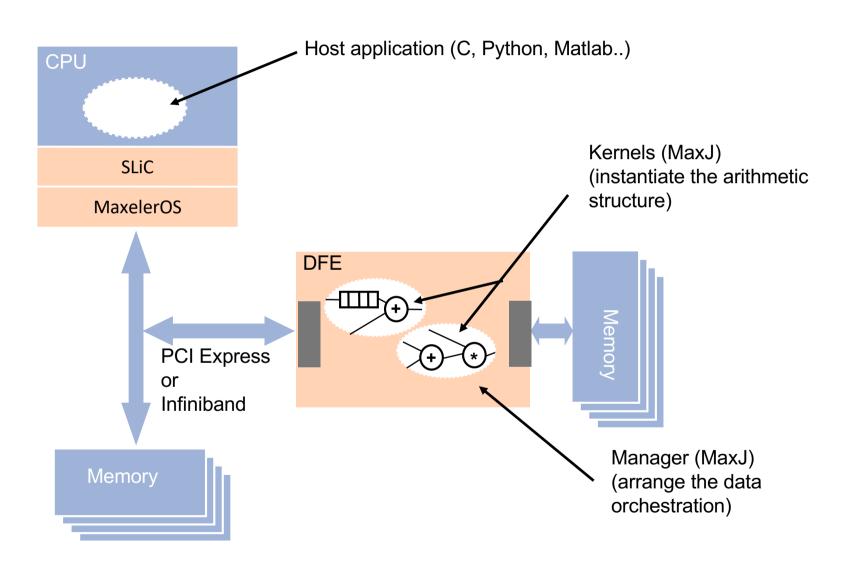


From Equations to Dataflow Hardware





Application Level Components





MaxRing

DataFlow Engine (DFE) Conceptual Model

LMEM

(Large Memory) 4-96GB

> High bandwidth memory link

Reconfigurable compute fabric

> Dataflow cores & **FMEM** (Fast Memory)

> > Link to main data network (e.g., PCIe, Infiniband)

- 48GB DDR3 DRAM (LMEM)
- Stratix V D8
- MaxRing interconnect
- 4,000 multipliers
- 700K logic cells
- 6.25MB of FMEM



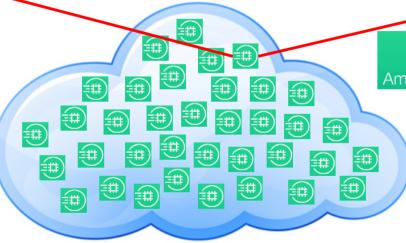
MAX4 (4-th generation) DFE



links

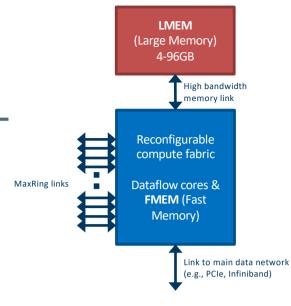
Maxeler's DFE (MAX5 gen)





Amazon EC2 F1 Instances

November 30, 2016, Las Vegas USA @ AWS re:Invent



Virtex UltraScale VU9P

- 1,182k Logic units
- 6,840 DSP blocks
- 43.3 MB FMEM

48GB LMEM

- 3 parallel SODIMM
- Low power DDR4

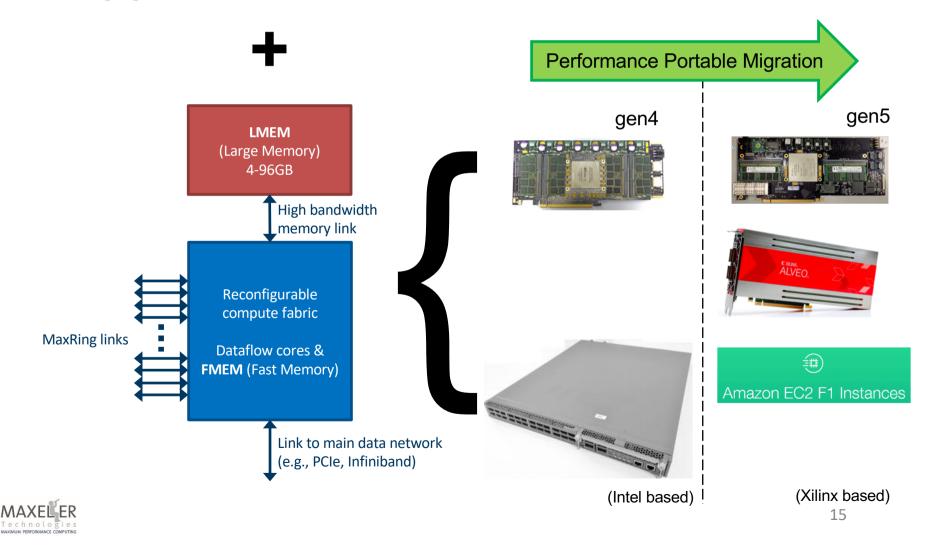




Fully compatible with Xilinx U200 and U250 Alveo

Multiple platforms, single DFE abstraction

Application and MaxJ



The Computational Model

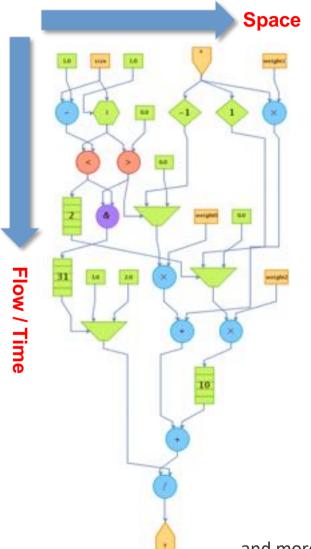
- Dataflow sub-system (DFE)
 - Spatial arithmetic chip "hardware" technology with flexible arithmetic units and programmable interconnect (looks like FPGAs but is not limited to)
 - Programmable Static Dataflow
 - Systolic Execution at kernel level
 - Streaming Custom Computing at system level
 - Implicit GALS* IO and kernel-to-kernel communication
- Dedicated software suite (MaxCompiler, MaxelerOS and SLiC)
 - compilation toolchain and design methodology
 - Incorporated simulation and debug environment for rapid development
 - Linux fully integrated runtime system and low level software support
 - Help designer focus on the data/algorithm and the system architecture
- Only three basic memory types (explicitly exposed)
 - Scalars (exposed to the CPU)
 - Fast Memory (FMEM): small and fast (on-chip)
 - Large Memory (LMEM): large and slow (off-chip)





^{*} GALS – Globally Asynchronous Locally Synchronous

Optimizations at all levels



Multiple scales of computing	Important features for optimization
complete system level	⇒ balance compute, storage and IO
parallel node level	⇒ maximize utilization of compute and interconnect
microarchitecture level	⇒ minimize data movement
arithmetic level	 ⇒ tradeoff range, precision and accuracy = discretize in Time, Space and Value
bit level	⇒ encode and add redundancy
transistor level	=> manipulate '0' and '1'

and more, e.g., trade/hide Communication (Time) for/behind Computation (Space)



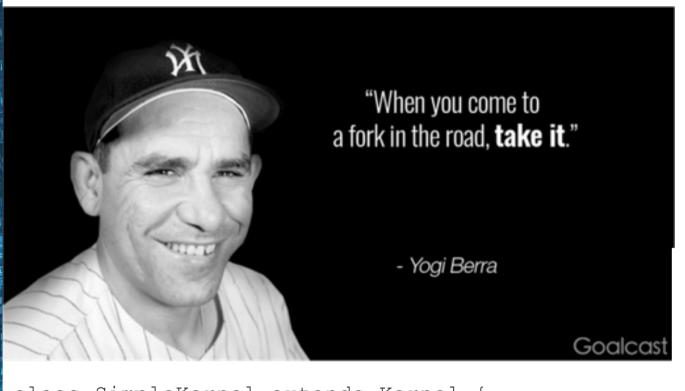
MaxJ: Moving Average of three numbers

Dataflow computing in hardware using a language you know

```
Similarle Search Broket Bun Window Strip-
     ** D 2 Novingheroprimpletered mad II
            B & T
                     class MovingAveKernel extends Kernel {
Y & lutorial chapit3 example3 movine
 It WE'CPU Code
 T project Code
                          MovingAveKernel(KernelParameters parameters) {
                                super(parameters);
    3 (2 MovingAverageSimple
     3 @ MovingAverageSimple
 w Aun Pules
                                DFEVar x = io.input("x", dfeFloat(8, 24));
  W A Simulation
     Build Log
                                DFEVar prev = stream.offset(x, -1);
     A Hanager Draph
                                DFEVar next = stream.offset(x, 1);
It has belonial chapit the example the
                                DFEVar sum = prev + x + next;
                                DFEVar result = sum / 3;
                                io.output("y", result, dfeFloat(8, 24));
                   Console II Simulated System Lag
                   Simulation (futurial chaptit) example) movingsveragesimple)
                                                                                                                                    MAXEEE
```

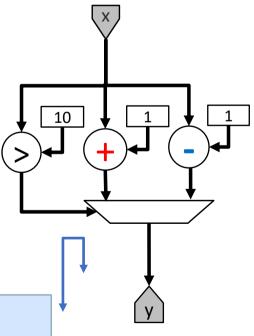


What about branches in Space



```
class SimpleKernel extends Kernel {
    SimpleKernel() {
        DFEVar x = io.input("x", dfeInt(24));
        DFEVar result = (x>10) ? x+1 : x-1;
        io.output("y", result, dfeInt(25));
    }
}
```



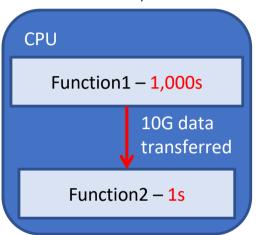


Maybe take both paths?

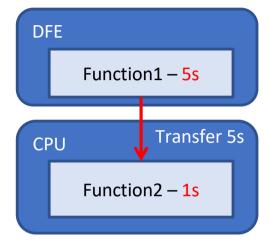


Decelerate to Accelerate

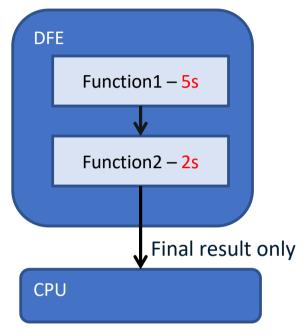
CPU time 1,001s



Option 1 time 11s



Option 2 time 7s



Some observations At Kernel level:

- Kernel 1 speedup
- Kernel 2 "speedup"

200x (!)

0.5x (!)

91x

143x

At System level:

- Option 1 (Kernel 1 only) speedup
- Option 2 (Kernels 1 and 2) speedup

But what about the required effort?



Easy it is not (and not really new)

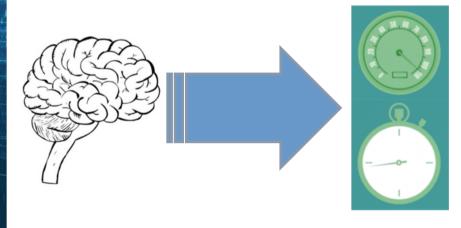
Slotnick's law (of effort):

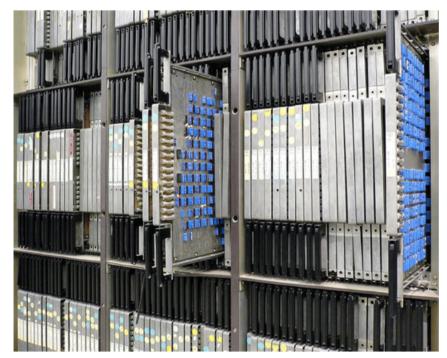
"The parallel approach to computing does require that some *original thinking* be done **about numerical analysis and data management** in order to secure efficient use.

In an environment which has represented the absence of the need to think as the highest virtue this is a decided disadvantage."



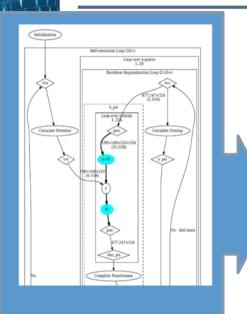
Daniel Slotnick (1931-1985) Chief Architect of Illiac IV



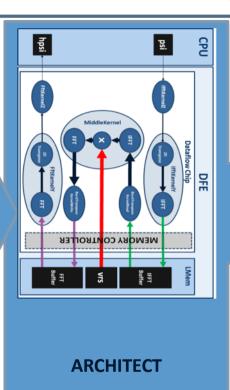


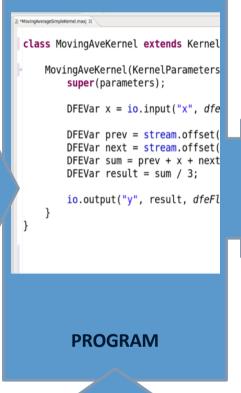


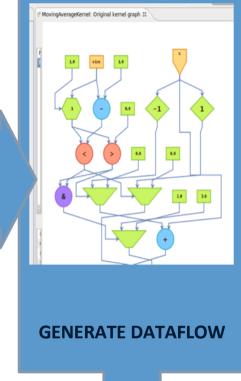
Non Traditional Design Process

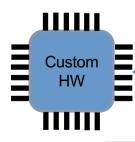


ANALYSE









many hours ...

OK?

SIMULATE AND DEBUG

Used to build **balanced** real systems, however, not easy to learn/educate





Open Research Questions

- How to compare against general purpose machines?
- How to validate results as "good enough"?
- How to estimate the required "Dan Slotnick's Effort"?
- How to significantly decrease P&R times?
- How to create a much better execution substrate?
- How to educate the think → model → program mindset?
- Tools, tools, tools, ...
 (and many, many more)

All of the above while dealing with growing impact of Quantum Mechanical Effects

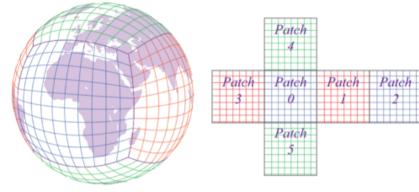




Global Weather Simulation with DFEs in China

An order of magnitude improvement over the Linpack-driven supercomputer technology

- ◆ L. Gan, H. Fu, W. Luk, C. Yang, W. Xue, X. Huang, Y. Zhang, and G. Yang, Accelerating solvers for global atmospheric equations through mixed-precision data flow engine, published at **FPL 2013**
- ◆ Joint research with Imperial College and Tsinghua University
- Simulating the atmosphere using the shallow water equation



(a) The cubed-sphere mesh

(b) The computational domain

Fig. 1. Mesh and computational domain.

Platform	Speedup	Energy Efficiency
6 Core CPU	1x	1x
Tianhe-1A Node	23x	15x
Maxeler MPC-X	330x	145x







But wait, there is someone who just knows





© Tom Philips





Backup



Why is all of this important?

'Tsunami of data' could consume one fifth of global electricity by 2025

Billions of internet-connected devices could produce 3.5% of global emissions within 10 years and 14% by 2040, according to new research, reports Climate Home News



▲ A Google data centre. US researchers expect power consumption to triple in the next five years as one billion more people come online in developing countries. Photograph: Google/Rex



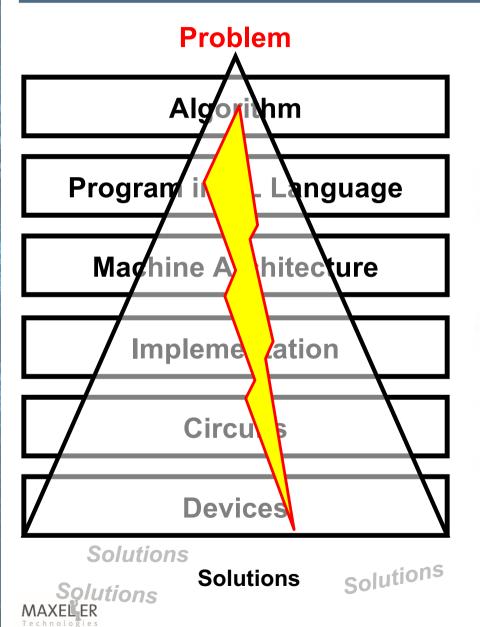
The Guardian

"... without dramatic increases in efficiency, ICT industry could use 20% of all electricity and emit up to 5.5% of the world's carbon emissions by 2025."

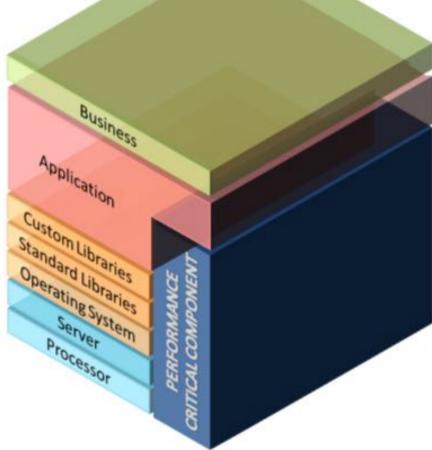
"We have a tsunami of data approaching. Everything which can be is being digitalised. It is a perfect storm."

"... a single \$1bn Apple data centre planned for Athenry in Co Galway, expects to eventually use 300MW of electricity, or over 8% of the national capacity and more than the daily entire usage of Dublin. It will require 144 large diesel generators as back up for when the wind does not blow."

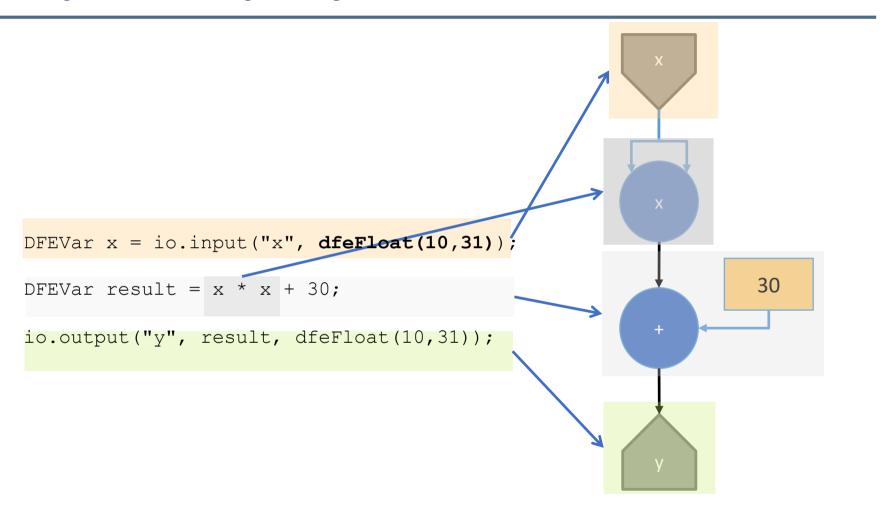
Solving Computing Problems Vertically



Co-optimise the HW and the SW stack for the performance critical areas of the application



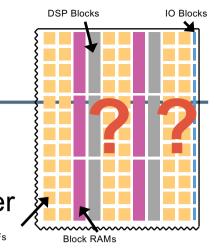
Simple example: $y = x^2 + 30$





DFE Resource Usage Reporting

- Allows you to see what lines of code are
- using what resources and focus optimization
- Separate reports for each kernel and for the manager



LUT/FFs

LUTs	FFs	BRAMs	DSPs	:	MyKernel.java		
727	871	1.0	2	:	resources used by this file	Different operations use different resources	
0.24%	0.15%	0.09%	0.10%	:	% of available		
71.41%	61.82%	100.00%	100.00%	:	% of total used		
94.29%	97.21%	100.00%	100.00%	:	% of user resources		
				:			
				: public class MyKernel extends Kernel {			
				:	<pre>public MyKernel (KernelParameters parameters) {</pre>		
				:	<pre>super(parameters);</pre>		
1	31		0				
2	9	0.0	0	:			
				:	<pre>DFEVar offset = io.scalarInput("offset", dfeUInt(8));</pre>		
8	8	0.0	0	:	DFEVar addr = offset + q ;		
18	40	1.0	0	:	<pre>DFEVar v = mem.romMapped("table", addr,</pre>		
				:	dfeFloat	2(8,24), 256);	
139	145	0.0	2	:	p = p * p;		
401	541	0.0	0	:	p = p + v;		
				:	<pre>io.output("r", p, dfeFloat(8,24));</pre>		
				:	}		
				:	}		



Optimization Feedback

 MaxCompiler gives detailed latency and area annotation back to the programmer

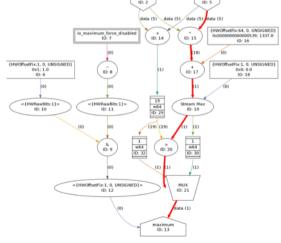
```
27
28
12.8:
d.Buy = ask.Price <= lowPrice & order_book.securityId === secId;
d.Sell = bid.Price >= highPrice & order_book.securityId === secId;
d.Quantity = d.Buy ? ask.Quantity : bid.Quantity;
d.Price = d.Buy ? ask.Price : bid.Price;

12.8ns

+ 6.4ns

19.2ns (total compute latency)
```

• Evaluate precise effect of code on latency and chip area





Pilot System Deployed at Jülich

Small pilot system deployed in Oct 2017

- one 1U MPC-X with 8 MAX5 DFEs
- one 1U AMD EPYC based server
- one 1U login head node

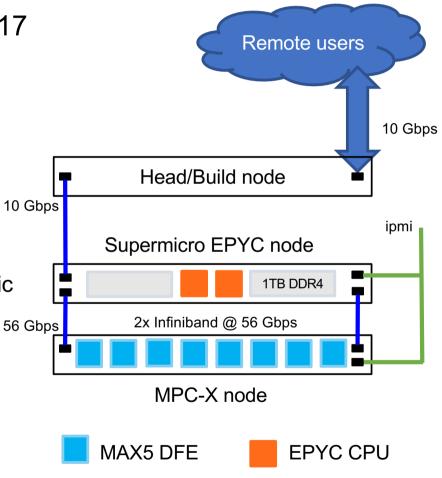
Scaling using Amazon AWS cloud

- MAX5 fully compatible with F1 instances
- Elastic scaling between private and public





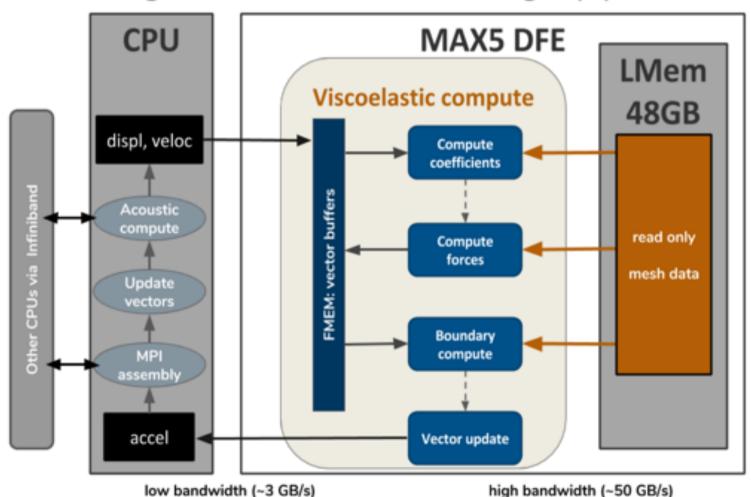
http://www.prace-ri.eu/pcp/





PRACE-PCP: SpecFEM3D on DFE

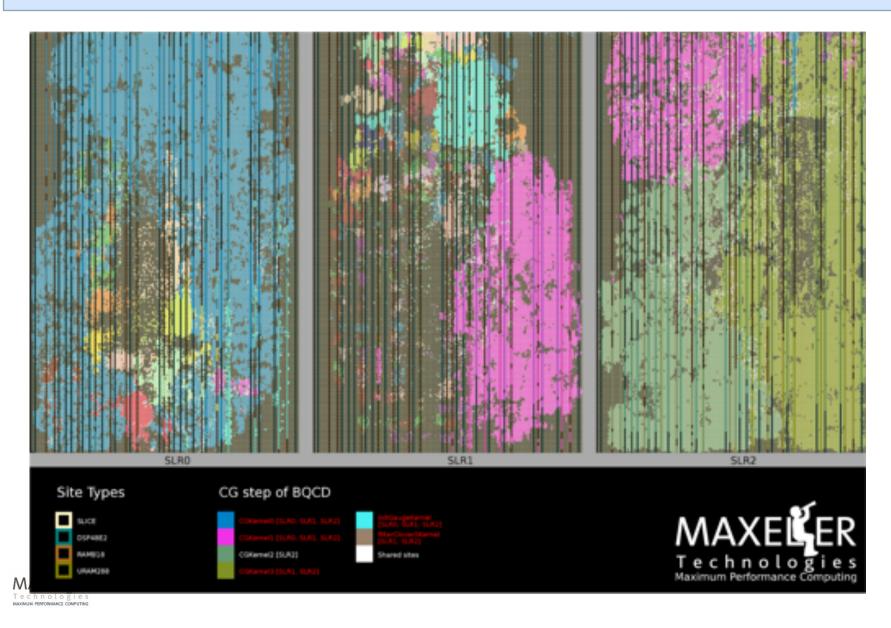
Heterogeneous architecture: single pipe



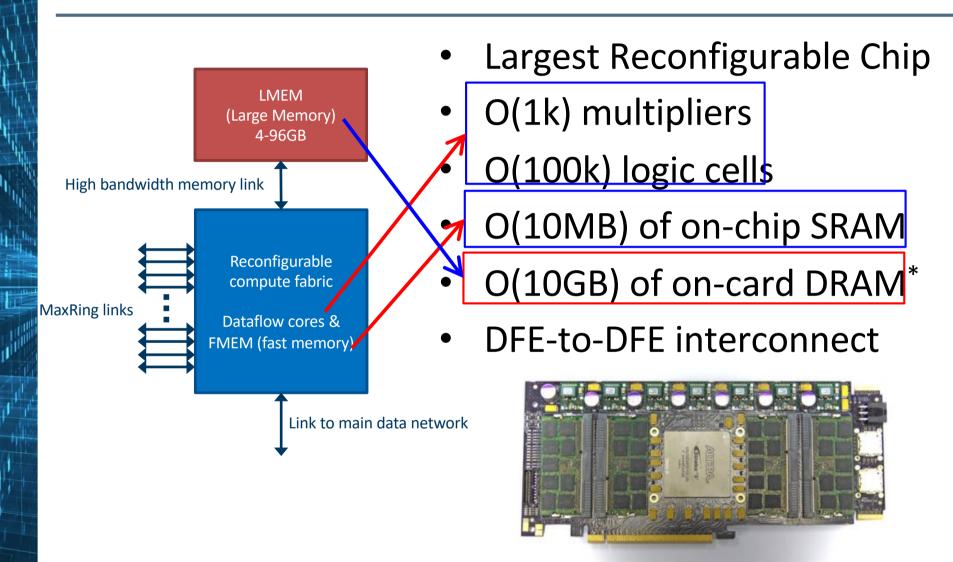


The BQCD Chip - AERIAL VIEW

Scalable Conjugate Gradient Design for the CG step of BQCD

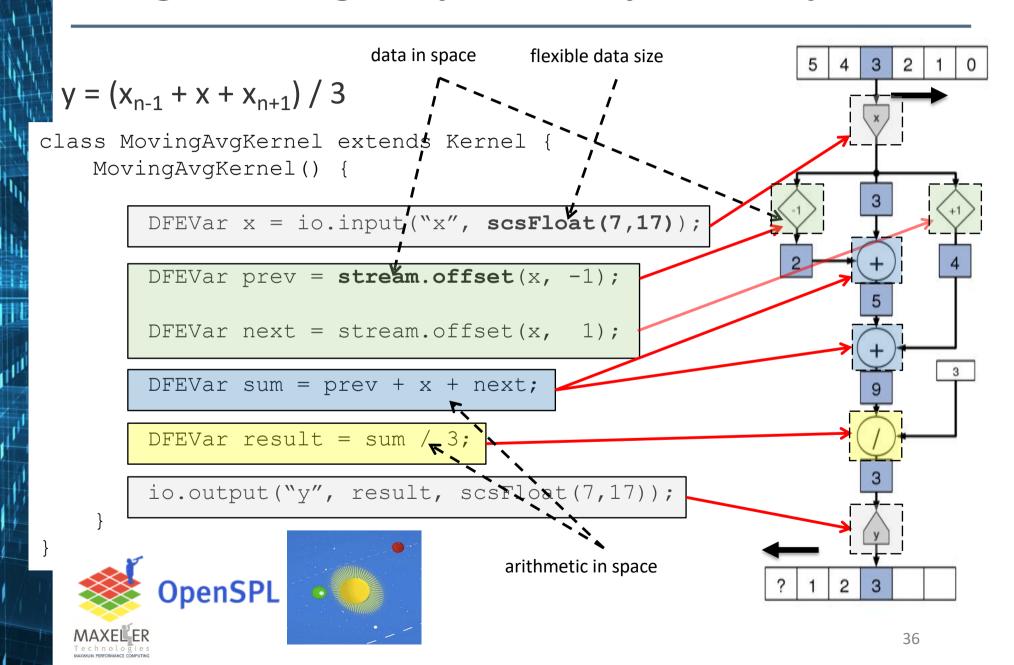


Maxeler's DataFlow Engines (DFEs)



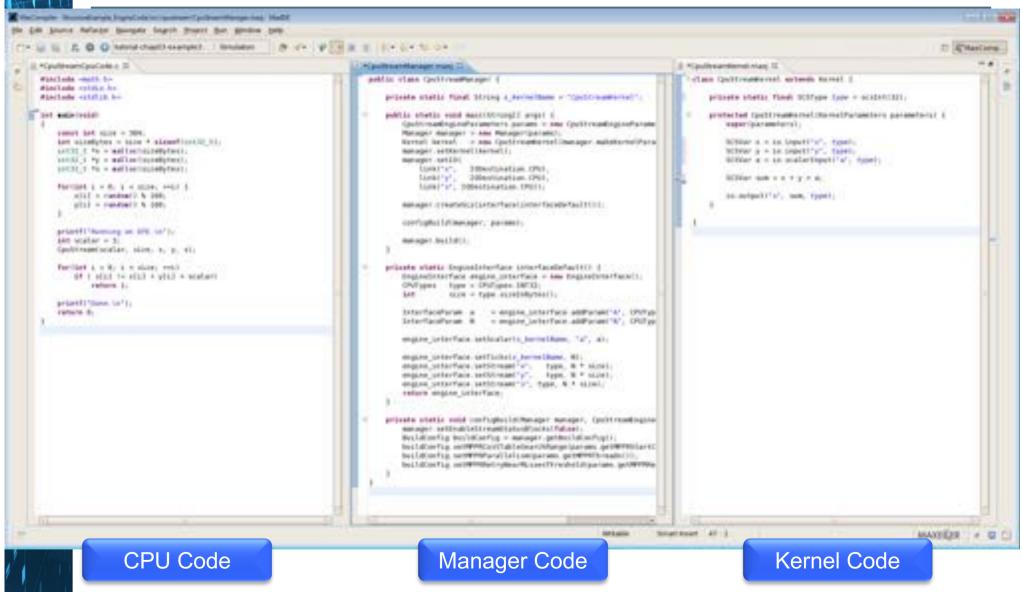
^{*} approaching 128GB on a ¾, single slot PCle card

Programming in Space: simple example





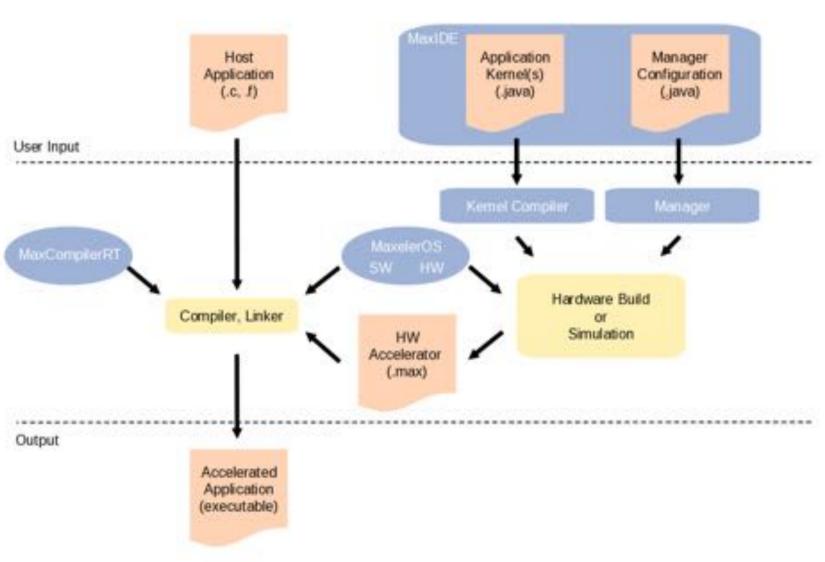
The required parts to create a DFE solution



MAXELER
Technologies
MAXIMUM PERFORMANCE COMPUTING

MAXIMUM PERFORMANCE COMPUTING

MaxCompiler Organization



Some links with more information

Maxeler Multiscale Dataflow Computing:

https://www.maxeler.com/technology/dataflow-computing/

Computing in Space explained by Mike Flynn:

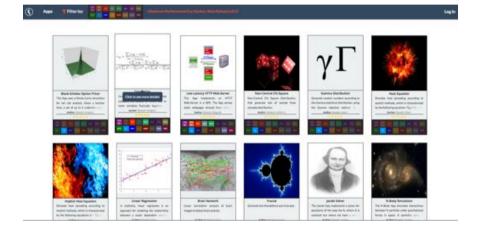
http://www.openspl.org/what-is-openspl/

Computing in Space Course at Imperial College:

http://cc.doc.ic.ac.uk/openspl16/

Exciting Applications for DFEs (and JDFEs):

http://appgallery.maxeler.com



Maxeler DFEs on AWS EC2 F1:

https://aws.amazon.com/marketplace/seller-profile?id=2780c6ec-d326-47fc-9ff6c66ab2ba202a

Maxeler and Xilinx Alveo collaboration:

https://www.xilinx.com/products/boards-and-kits/alveo.html

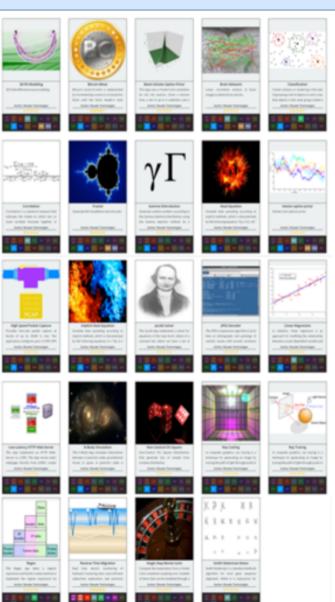




Maxeler Applications Gallery

Dataflow Apps and Analytics for Machine Learning

http://appgallery.maxeler.com/



Dataflow Engine (DFE) Ecosystem

- With over 150 universities in our university program, we decided to create an app gallery to enable the community to share applications, examples, demos, ...
- ◆ The App Gallery is complemented by a teaching program, with the first successful course taught at Imperial College in 2014, see

http://cc.doc.ic.ac.uk/openspl14

- ◆ Top 10 APPS:
 - > Correlation: in real-time, pairwise, on 6,000 streams
 - 100% Guaranteed Packet Capture
 - Webserver, cache and load balancing
 - HESTON Option pricer
 - N-body simulation
 - Regex matching (e.g. for Security)
 - Brain network simulation
 - Quantum Chromo-Dynamics kernel
 - Seismic Imaging
 - Realtime Classification

Over 150 Maxeler University Program Members





Universität Hamburg DER FORSCHUNG I DER LEHRE I DER BILDUNG



























































































