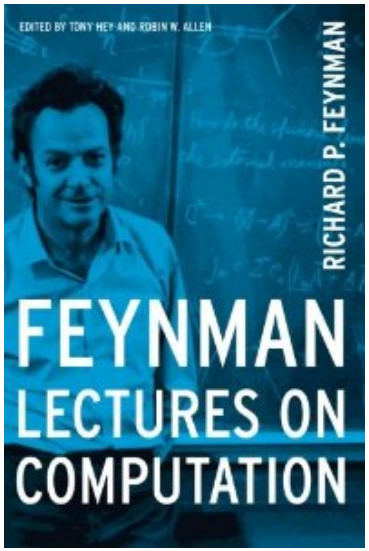
Multiscale Dataflow Computing

Oskar Mencer, HPSP Malta 2014

MAXELER
Technologies
MAXIMUM PERFORMANCE COMPUTING

Euclids Elements, Representing $a^2+b^2=c^2$

Richard Feynman on Computation



In theory, a computer system
can be constructed which uses no energy.

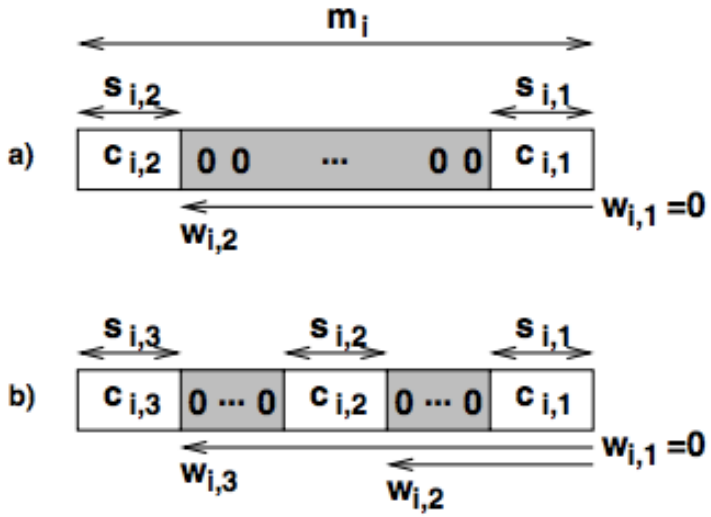
Energy is only needed when **information** is lost.

Reordering of **information** does not require energy
from a pure physics perspective.

Of course, moving **information** takes Energy…

so whats the most efficient technology to move data?

Expl: Minimize ‘1’s => Sparse Coefficients



$$p = \frac{32799}{32768} - \frac{609}{32768}x - \frac{14881}{32768}x^2.$$

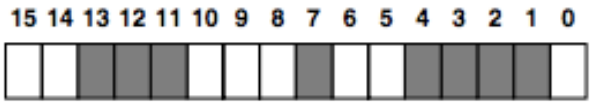


Fig. 2. Target format for cos function.

$$p = -\frac{75}{32768} + \frac{34538}{32768}x - \frac{6169}{32768}x^2.$$

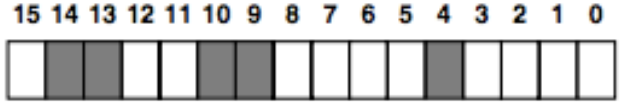


Fig. 3. Target format for sin function.

$$p = \frac{32793}{32768} + \frac{31836}{32768}x + \frac{21146}{32768}x^2.$$

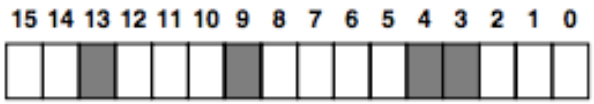
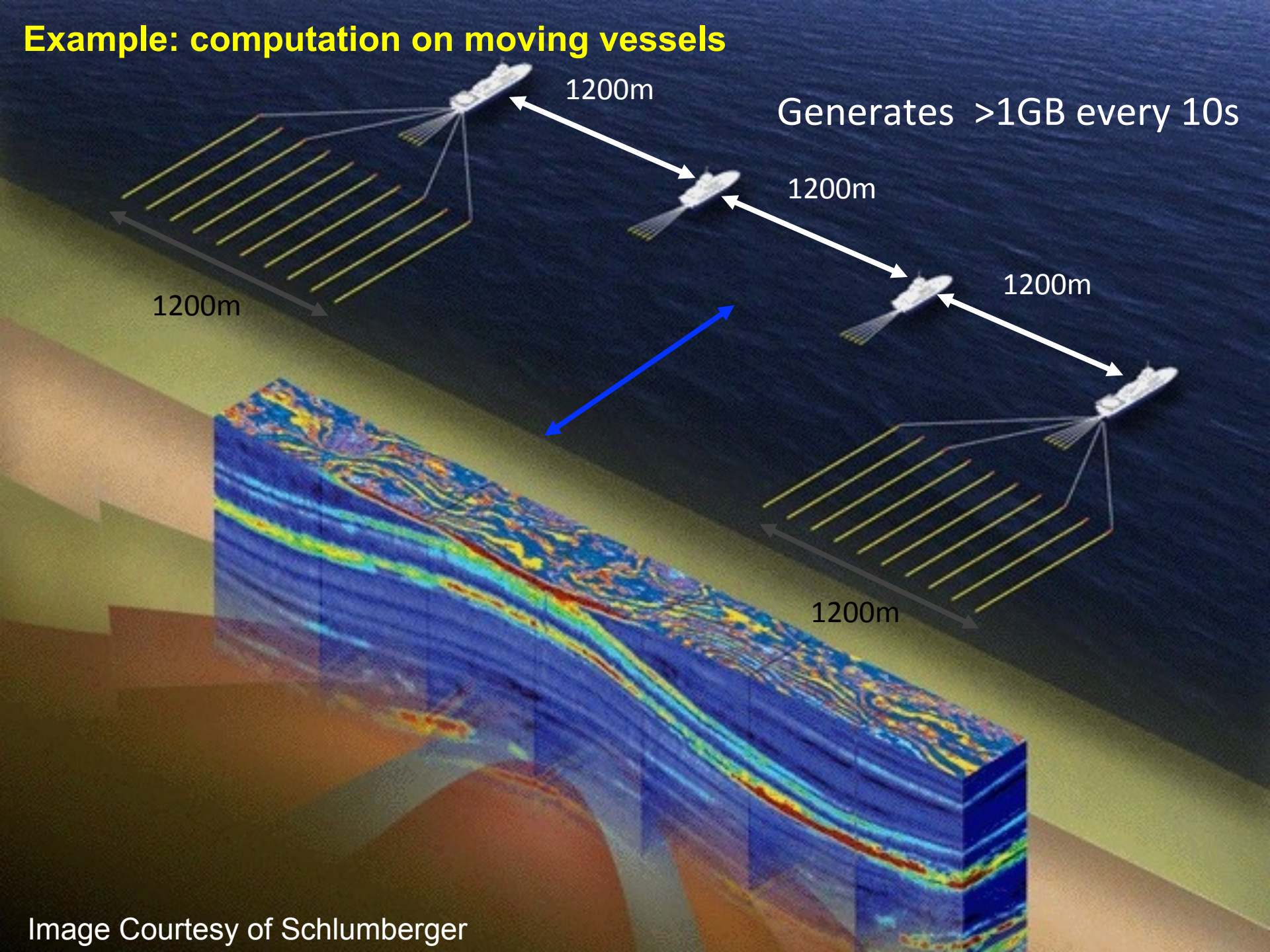


Fig. 4. Target format for exp function.

Nicolas Brisebarre, Jean-Michel Muller and Arnaud Tisserand
Sparse Coefficient Polynomial Approximations for Hardware Implementation,
Asilomar Conference, 2004.

Example: computation on moving vessels



Seismic Surface-related Multiple Elimination

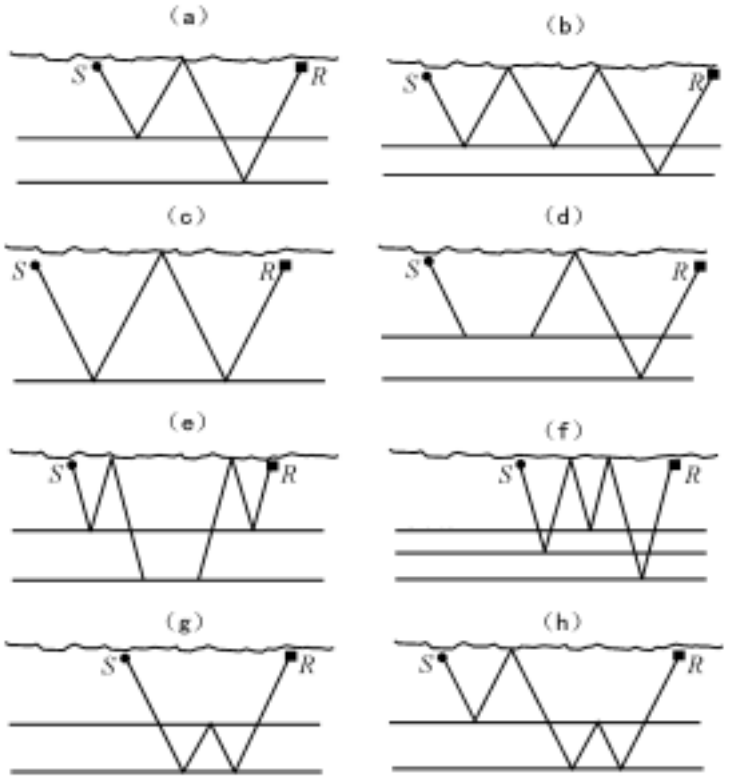
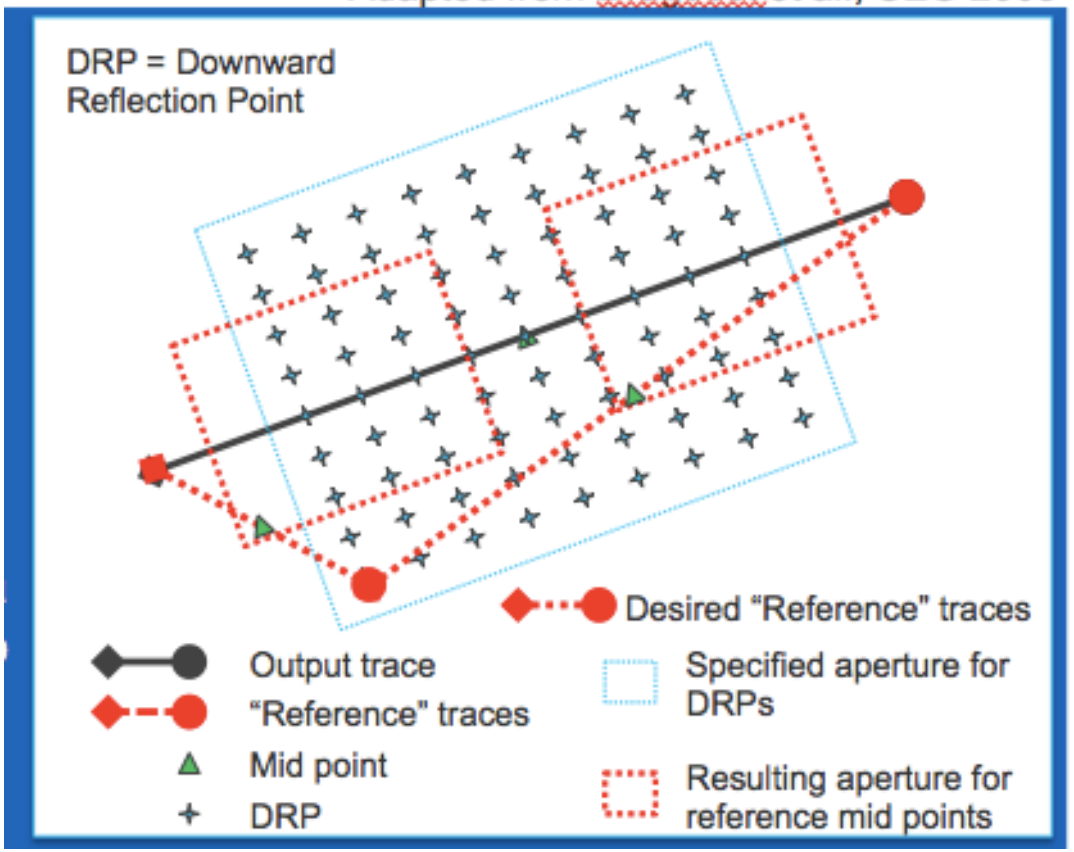


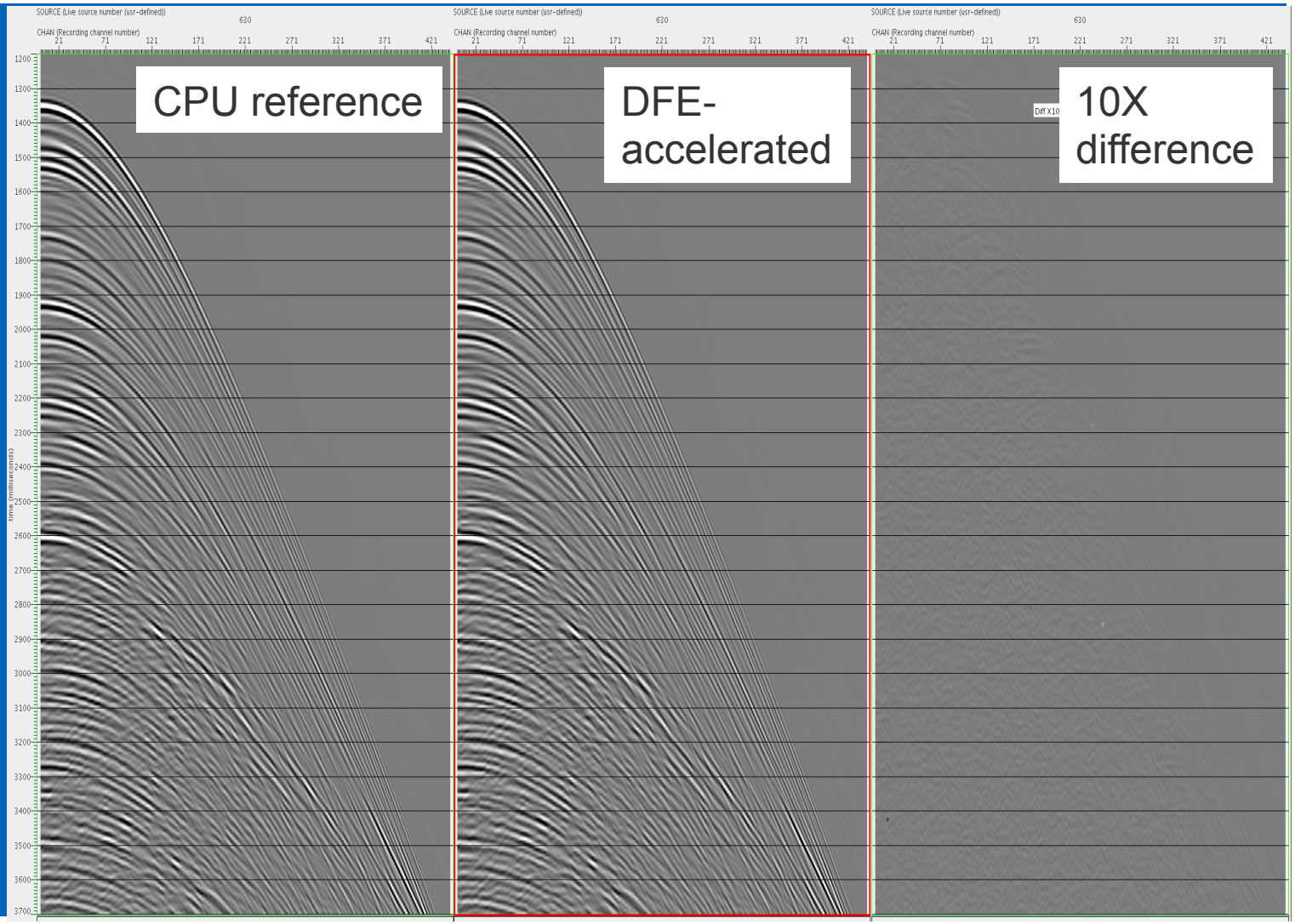
Fig. 1 Schematic diagram for raypaths of some typical multiple events

Huang et al.

Adapted from Dragoset et al., SEG 2008

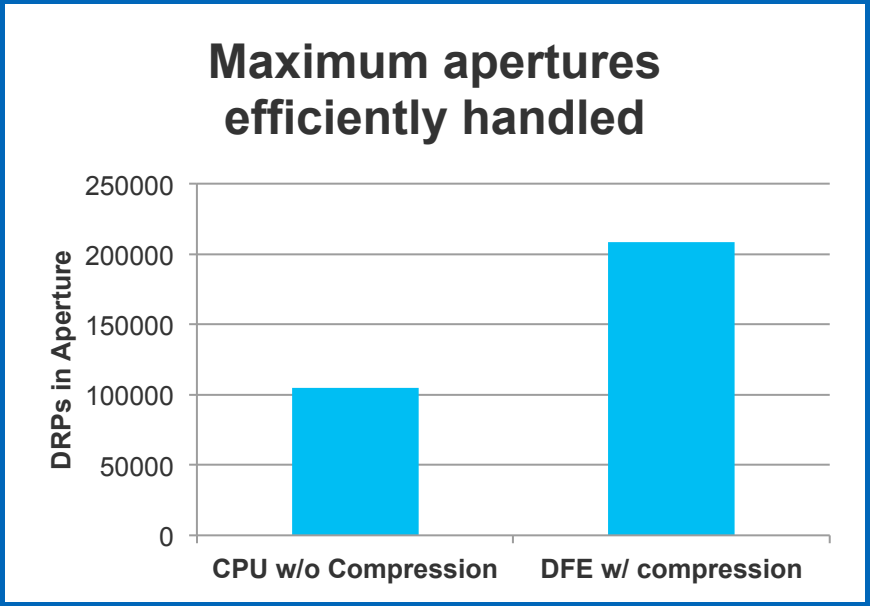


Comparing DFE output with CPU implementation
Compression with 30dB SNR



SRME results based on data from a 3D NATS dataset

- Comparing
  - 1U, 16-core, 2.7MHz, 128 GB, Sandy-Bridge E5-2680
  - 1U, 8 DFE, MPCX-2000
- The comparison is only for the accelerated “NMO-Convolution-chain”



NMO-Convolution Chain	Mpairs/s	Speed-up
16-core CPU	0.159	1X
MPCX-2000	6.4	40X

Maxeler Data Flow Engines - DFEs - 2014

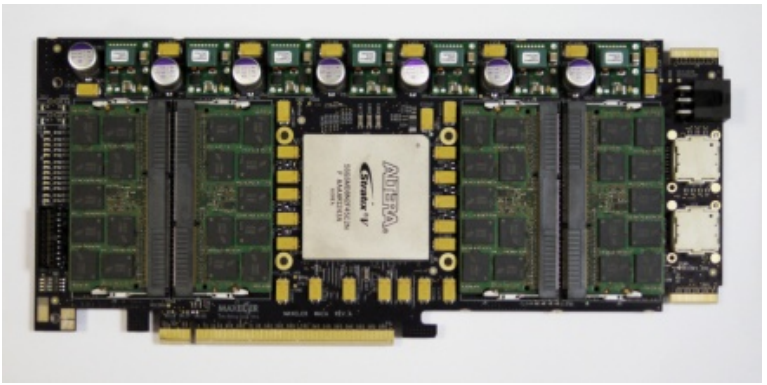
Juniper / Maxeler Ikon Switch

- 28nm process
- 250MHz clock frequency
- Sub 1us latency
- 16K TCP connections

MAX4N

- 28nm process
- 250MHz clock
- Sub 1us latency
- 16K TCP connections

MAX4



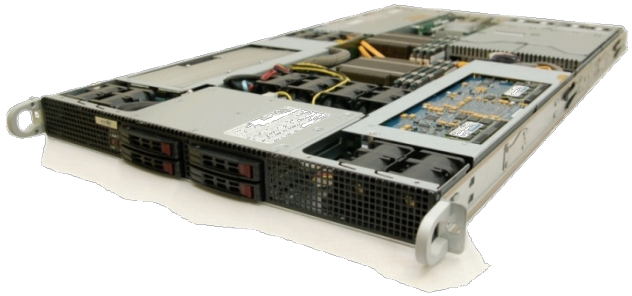
- 28nm process
- 250MHz clock frequency
- 6.25MB Fmem
- 4,000 multipliers
- 700K logic cells
- 3GB/s CPU bandwidth
- 384GB DRAM per 1U

Maxeler Dataflow Engines (DFEs)

MPC C Series

High Density DFEs

Intel Xeon CPU cores and up to 6 DFEs with 576GB of RAM



MPC X Series

The Dataflow Appliance

Dense compute with 8 DFEs, 768GB of RAM and dynamic allocation of DFEs to CPU servers with zero-copy RDMA access

MPC N Series

The Low Latency Appliance

Intel Xeon CPUs and 1-2 DFEs with direct links to up to six 10Gbit Ethernet connections

MaxWorkstation

Desktop dataflow development system

MaxRack

10, 20 or 40 node rack systems integrating compute, networking & storage

MaxCloud

Hosted, on-demand, scalable accelerated compute

Dataflow Engines

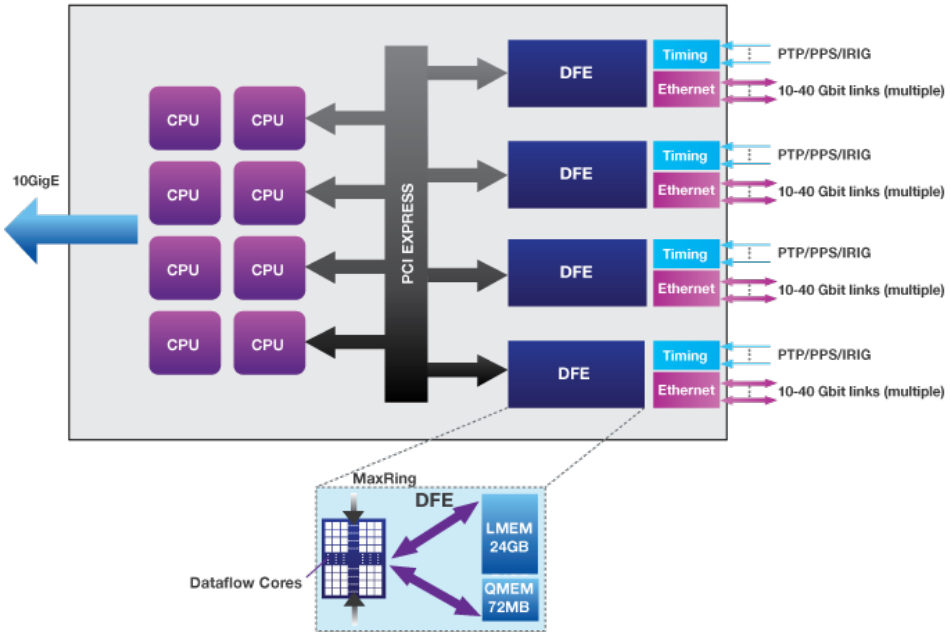
48GB DDR3, high-speed connectivity and dense configurable logic

MAXELER Technologies

Maxeler Dataflow Networking Platform: MPC-N32x

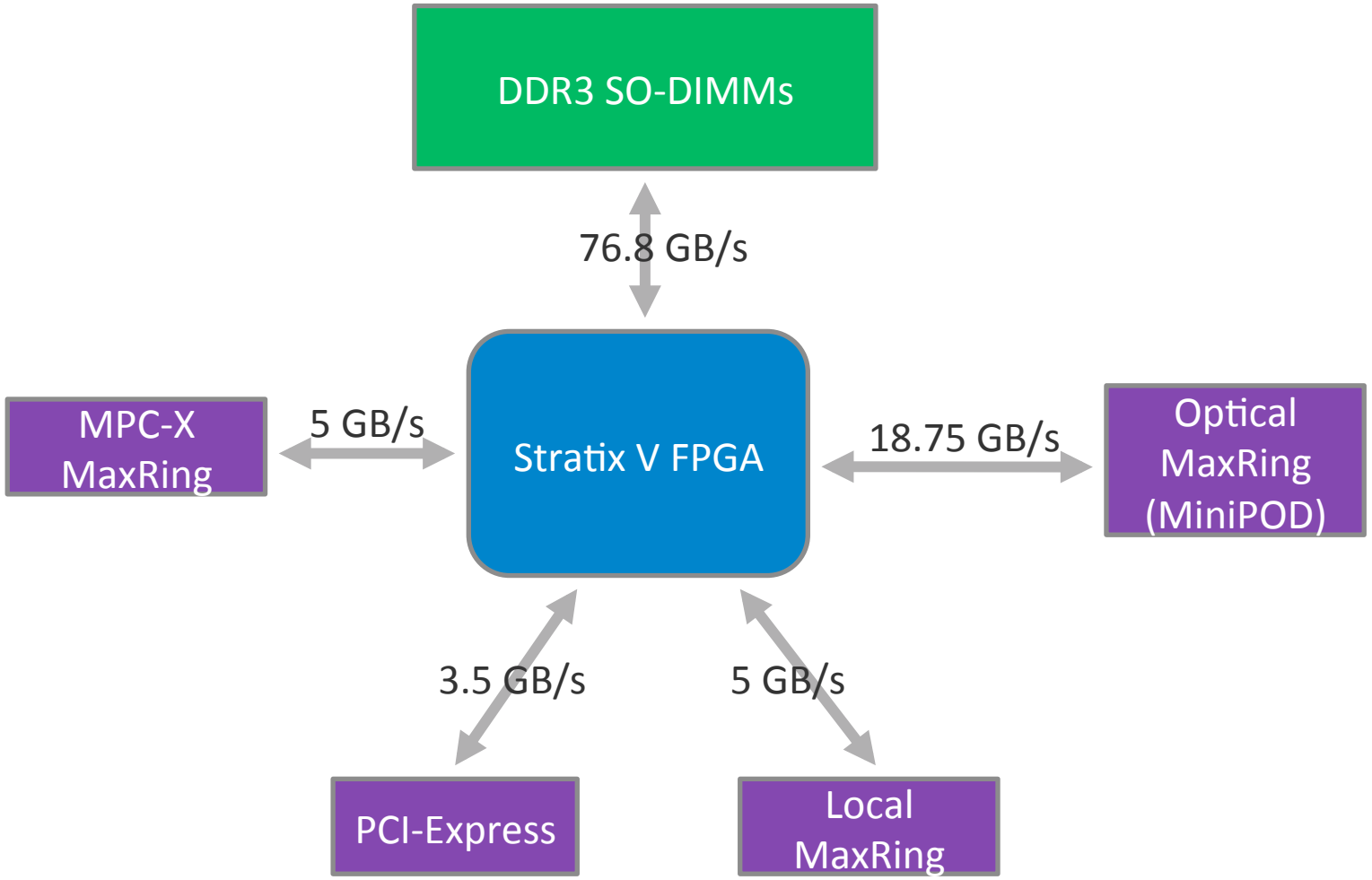
- Direct network connections to *isca* class dataflow engines (DFEs)
- Density: Up to 4 DFEs per 1RU
- Network Links per DFE
  - 1,2 or 3 zQSFP+ @ 40Gbps (or 4x10Gbps)
  - *3 x 100Gbps coming Q4-2013*
- Wire-to-wire latency 500ns
- Memory per DFE
  - 6.5MB SRAM @ 14.4TB/s
  - 72MB QDR @ 9.9GB/s
  - 24GB DRAM @ 38.4GB/s

MPC-N32x Series Architecture

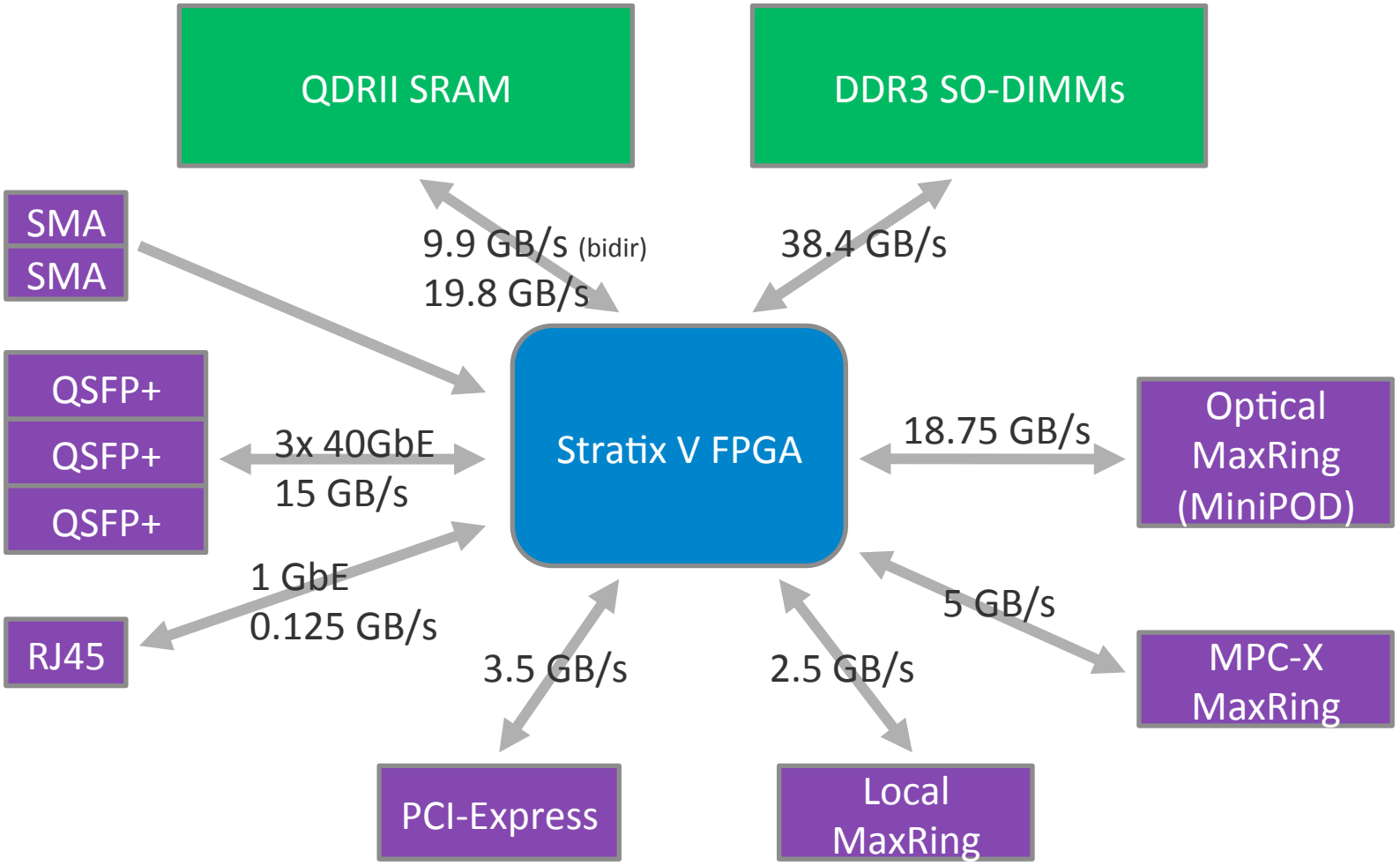


MAXELER Technologies

Maia DFE connectivity



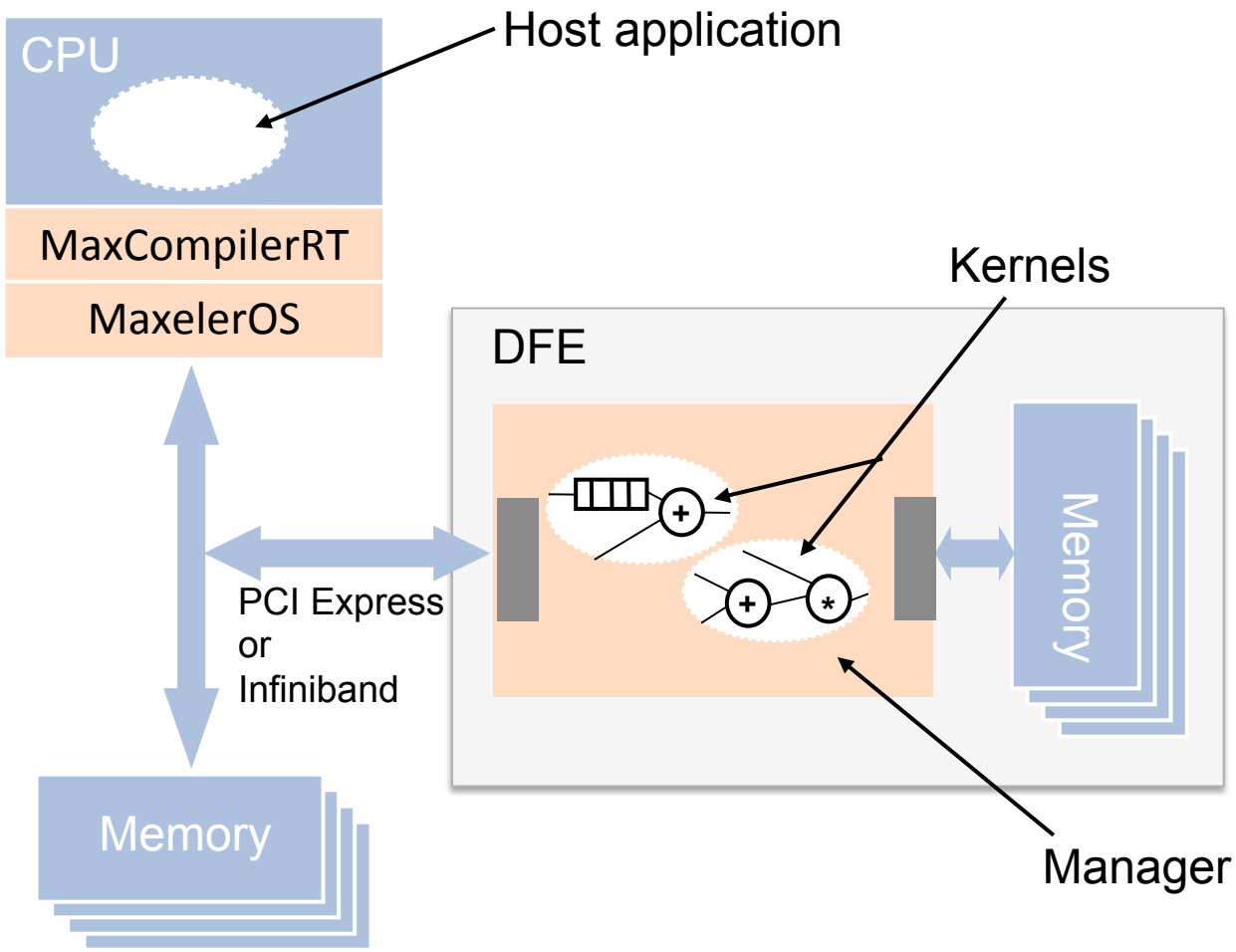
Isca DFE connectivity



MAXELER Technologies

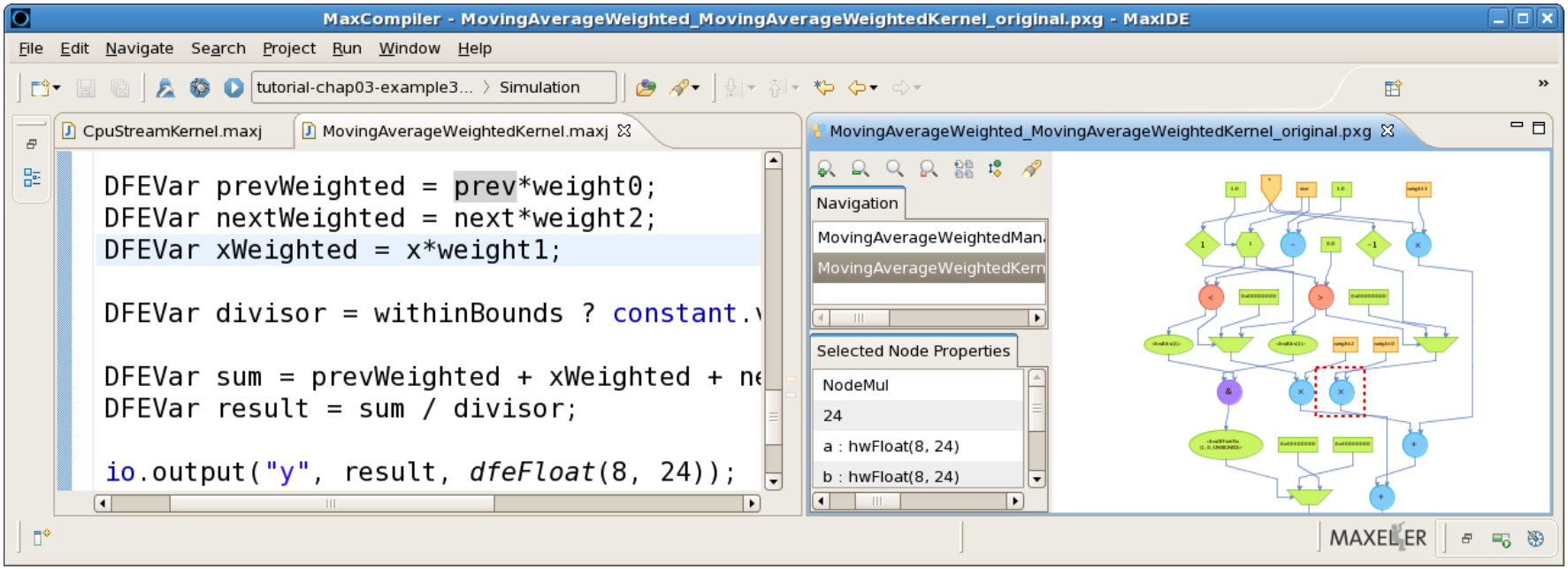
Maxeler DFE System Infrastructure

(developed over 20 years)



Maxeler VHDL generation from Java

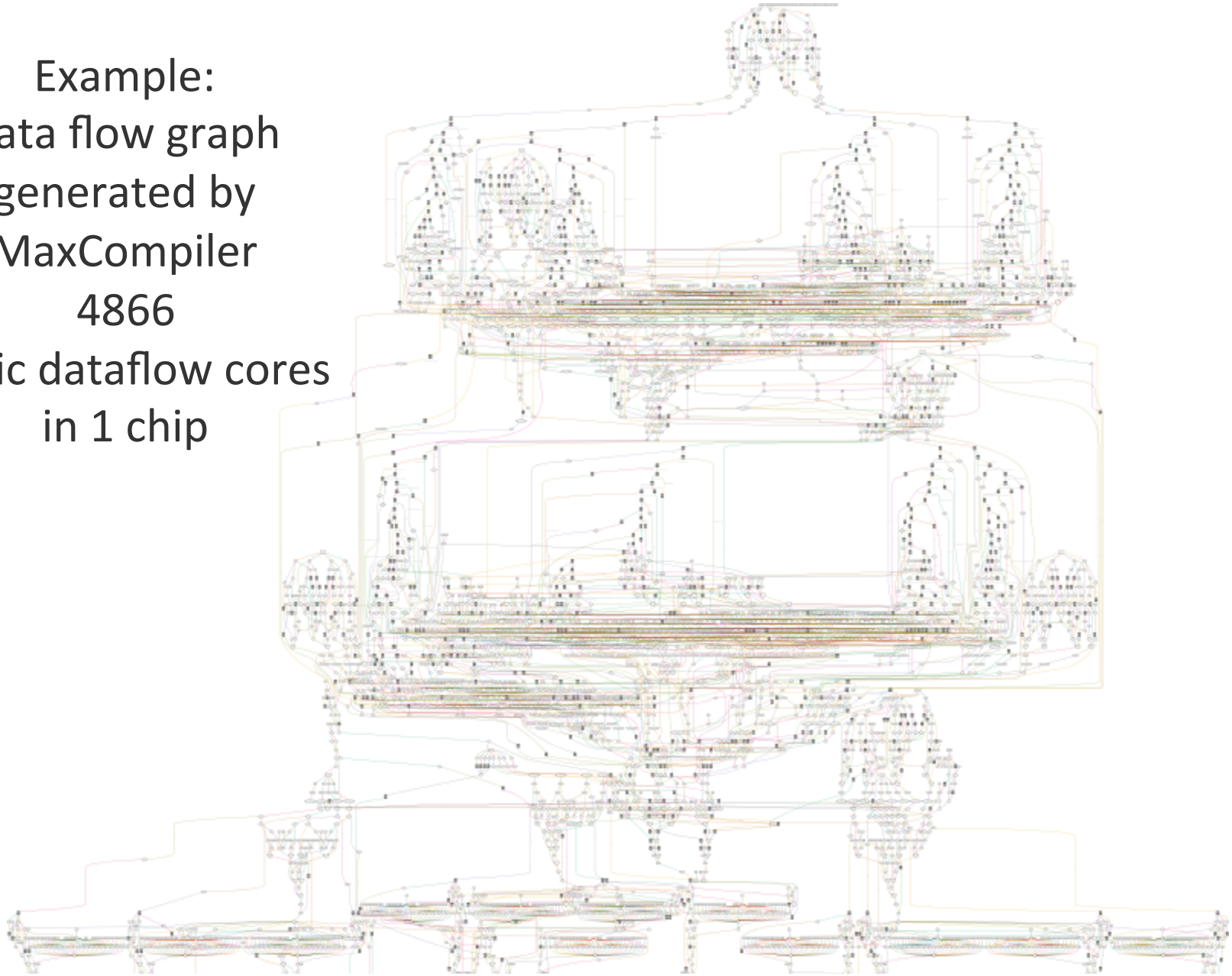
- **MaxCompiler** – Java-driven dataflow compiler creating near optimal VHDL

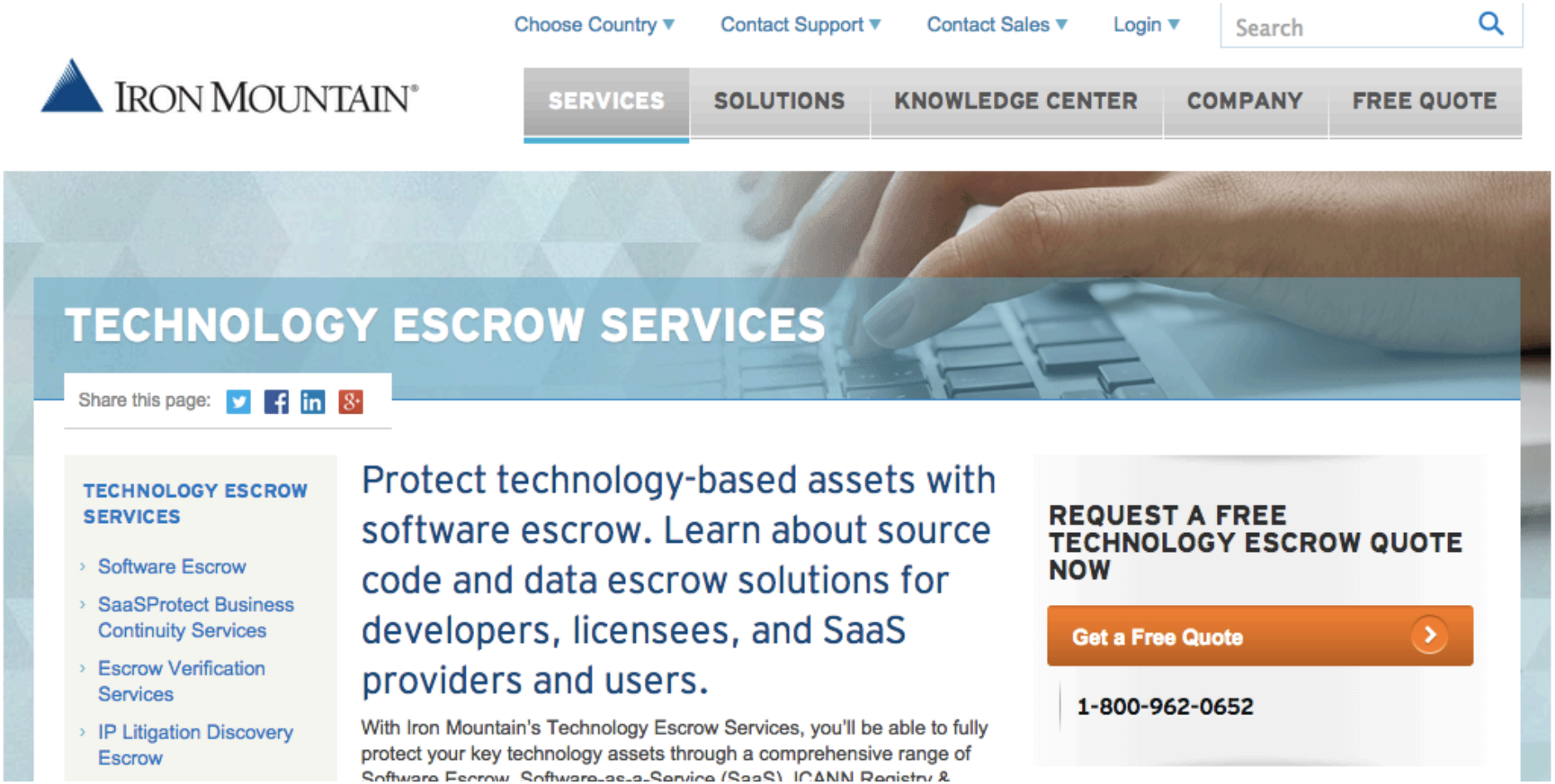


Making FPGA programming fun and easy, close to the sensor as well as for backend processing in the datacenter.

Making FPGAs programmable for high throughput as well as very complex computations!

Example:
data flow graph
generated by
MaxCompiler
4866
static dataflow cores
in 1 chip





Case Study – UK Government

Industry

Government

Engagement type

Hardware and Software Platform Sale

Main Contacts

Ministry of Science

Date

Dec 2013

Critical Client Issues

- Competitive advantage in international race
- Transitioning to Big Data Analytics while conventional solutions do not manage to keep up
- High Energy Physics keeps pushing computational demands

Approach

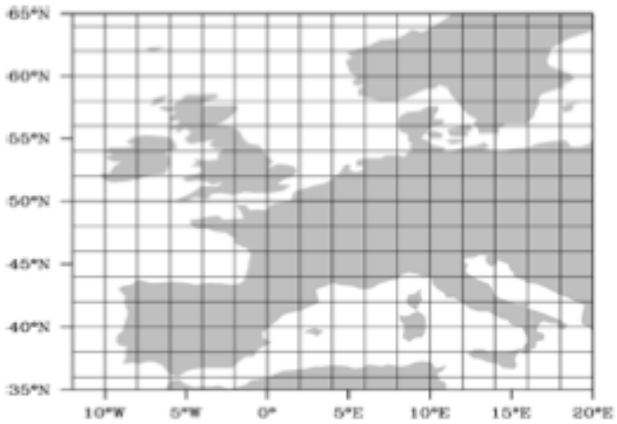
The approach:

1. Selecting Multiscale Dataflow Computing architecture
2. Selecting Maxeler Software Platforms
3. Training of staff to work with Maxeler technology
4. Support and maintenance of the installation

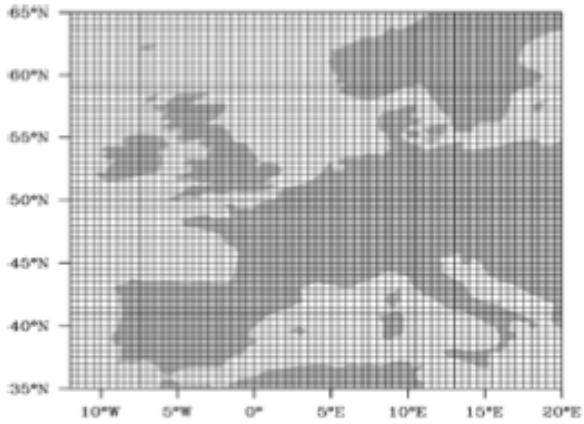
Client benefits

- 20-50x increased compute capability per cubic-foot of data center space
  => single Maxeler rack brings compute capability of over 20 conventional racks
- Enabling the evaluation of portable Petascale computing systems
- Green computing: chance to beat the top machines in the Green500 supercomputer list

Weather and climate models



Which one is better?



Finer grid and higher precision are obviously preferred but the computational requirements will increase ➔ Power usage → $$

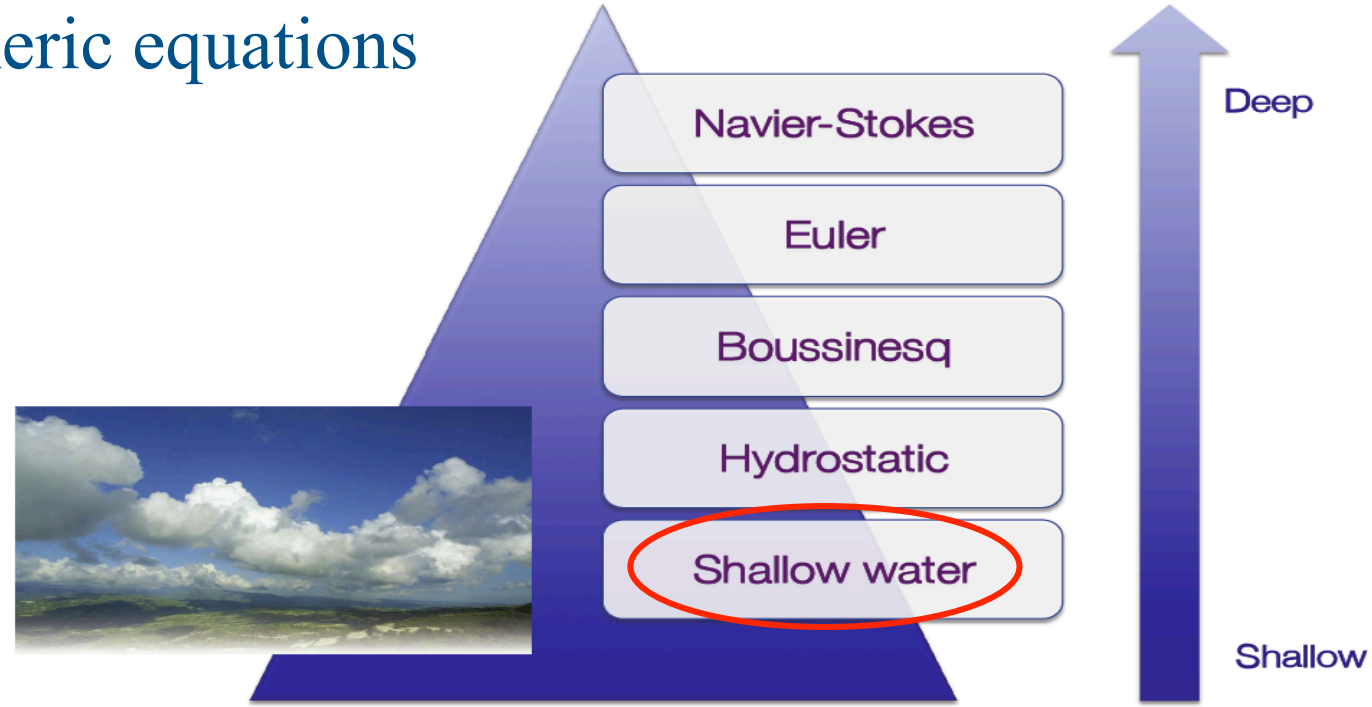
What about using reduced precision? (15 bits instead of 64 double precision FP)

We use only **15 bits** for 98% of the computation:

MAXELER Technologies

Global Weather Simulation

- Atmospheric equations



- Equations: Shallow Water Equations (SWEs)

[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, FPL2013]

Weather model -- performance gain

Platform	Performance ()	Speedup
6-core CPU	4.66K	1
Tianhe-1A node	110.38K	23x
MaxWorkstation	468.1K	**100x**
MaxNode	1.54M	**330x**

Meshsize: **14x**

MaxNode speedup over Tianhe node: 14 times

Imperial College London

ISCAS

MAXELER Technologies

Weather model -- power efficiency

Platform	Efficiency ( )	Speedup
6-core CPU	20.71	1
Tianhe-1A node	306.6	14.8x
MaxWorkstation	2.52K	**121.6x**
MaxNode	3K	**144.9x**

9 x

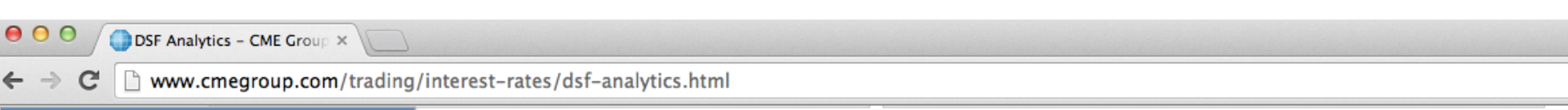
Meshsize:

MaxNode is 9 times more power efficient

Imperial College London

ISCAS

MAXELER Technologies



Powering Exchange Analytics at the Chicago Mercantile Exchange CME Group, 2014

onal use only using Maxeler Technologies' curve construction methodology. This tool uses delayed data and displayed results are indicative representations only.

Please hover your mouse pointer over column titles and links for further information.

CME Ticker	Bloomberg Ticker	DSF Pricing					Timestamp
		Price	Coupon	PV01	NPV	Implied Rate	
T1UM4 2Y	CTPM4	100'057	0.750%	$19.97	$179.69	0.6600%	4:00:03 PM CT 4/4/2014
F1UM4 5Y	CFPM4	100'115	2.000%	$48.49	$359.38	1.9259%	4:00:03 PM CT 4/4/2014
N1UM4 10Y	CNPM4	100'225	3.000%	$90.16	$703.12	2.9220%	4:00:03 PM CT 4/4/2014
B1UM4 30Y	CBPM4	102'270	3.750%	$195.07	$2,843.75	3.6042%	4:00:03 PM CT 4/4/2014
T1UU4 2Y	CTPU4	100'085	1.000%	$19.93	$265.62	0.8668%	4:00:03 PM CT 4/4/2014
F1UU4 5Y	CFPU4	100'110	2.250%	$48.27	$343.75	2.1788%	4:00:03 PM CT 4/4/2014
N1UU4 10Y	CNPU4	101'125	3.250%	$89.55	$1,390.62	3.0948%	4:00:03 PM CT 4/4/2014
B1UU4 30Y	CBPU4	106'020	4.000%	$193.47	$6,062.50	3.6868%	4:00:03 PM CT 4/4/2014

Quotes and analytics are updated every 15 minutes.

Analytics powered by Maxeler Technologies©

Instrument	CPU 1U-Node	Max 1U-Node	Comparison
European Swaptions	848,000	35,544,000	42x
American Options	38,400,000	720,000,000	19x
European Options	32,000,000	7,080,000,000	221x
Bermudan Swaptions	296	6,666	23x
Vanilla Swaps	176,000	32,800,000	186x
CDS	432,000	13,904,000	32x
CDS Bootstrap	14,000	872,000	62x

American Finance Technology Award with JP Morgan, Credit Derivatives Risk, 2011

MAXELER Technologies

Example: Quad Precision Floating Point

Quad Precision (112 bit mantissa, 15 bit exponent)
One Maxeler Maia DFE = 21.4 GFLOP/s
1U MPC-X2000 with 8 Maias = 171.2 GFLOP/s.

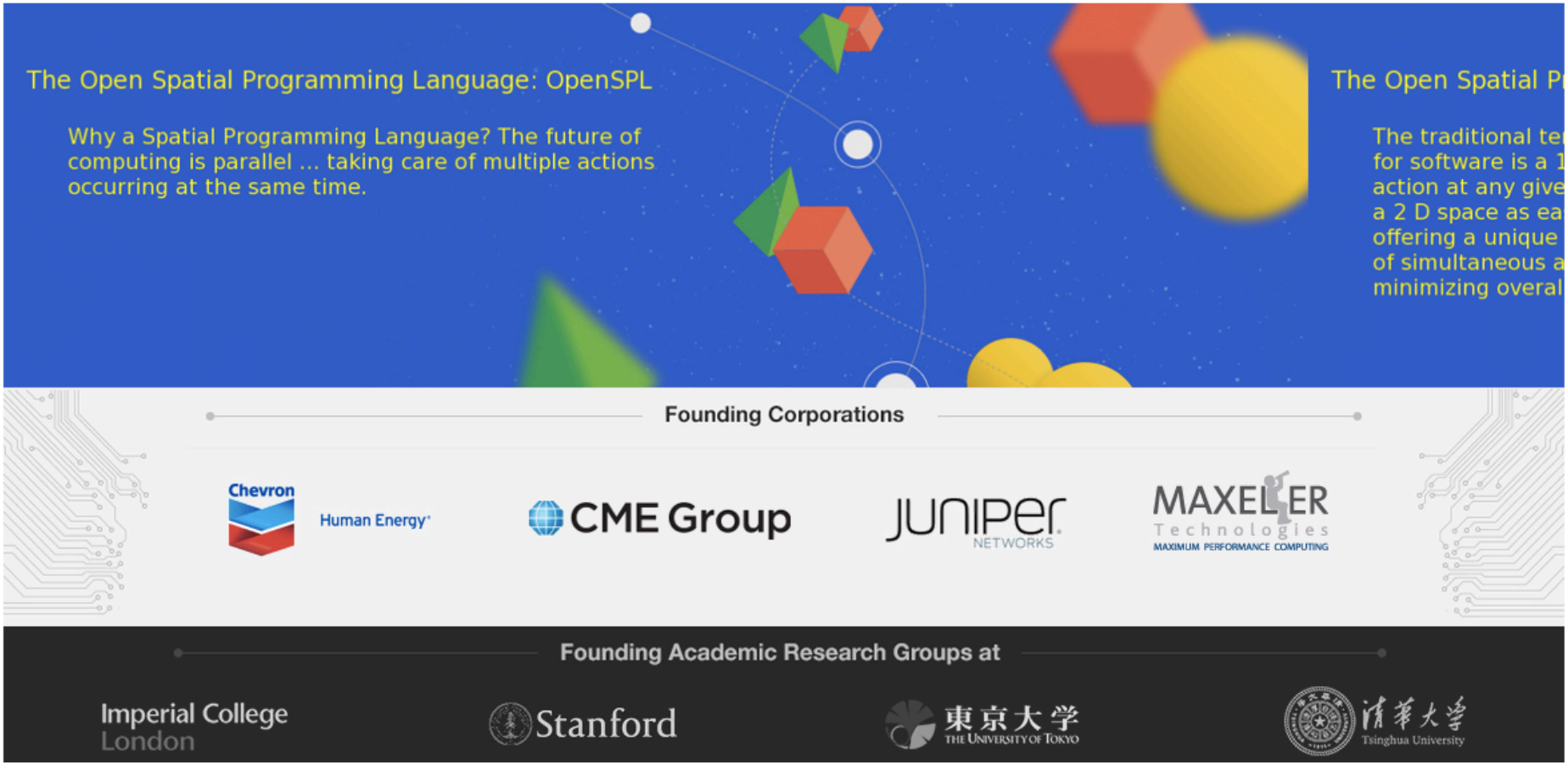
1U server node of Sandybridge 16-core is estimated to run at 1.05 GFLOP/s.

1U to 1U: ~160x.

OpenSPL Foundation

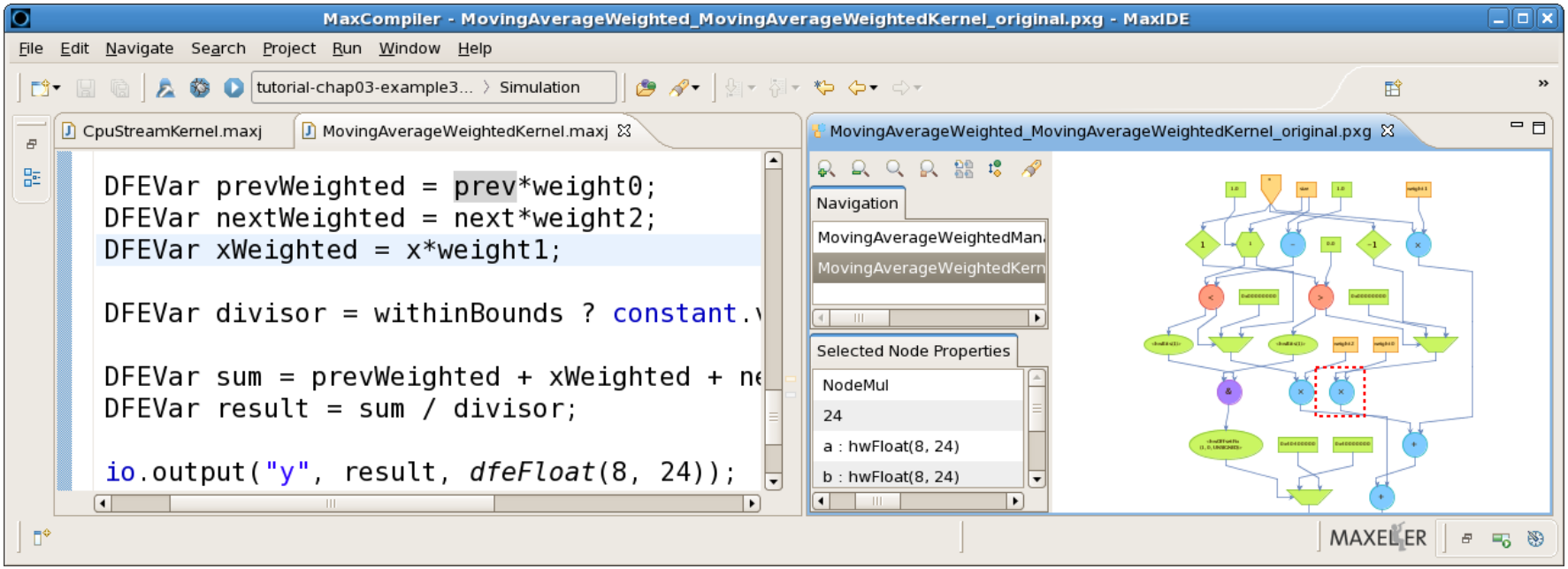
"OpenSPL enables us to build parallelized applications that fully take advantage of spatial computing technology with the ease of a high-level software project"

– Ryan Eavy, Executive Director, Architecture, CME Group



Maxeler DFE Programming

- **MaxCompiler** – Java-driven dataflow compiler



Making FPGA programming fun and easy, close to the sensor as well as for backend processing in the datacenter.

Making FPGAs programmable for high throughput as well as very complex computations!

Chip Resource Usage

The goal is to maximize utilization of resources on the chip, and bandwidth on the memory bus.

```
  LUTs      FFs    BRAMs     DSPs : MyKernel.java
   727      871      1.0        2 : resources used by this file
 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 {
                                 :   public MyKernel (KernelParameters parameters) {
                                 :     super(parameters);
     1       31      0.0        0 :     DFEVar p = io.input("p", dfeFloat(8,24));
     2        9      0.0        0 :     DFEVar q = io.input("q", dfeUInt(8));
                                 :     DFEVar offset = io.scalarInput("offset", dfeUInt(8));
     8        8      0.0        0 :     DFEVar addr = offset + q;
    18       40      1.0        0 :     DFEVar v = mem.romMapped("table", addr,
                                 :                              dfeFloat(8,24), 256);
   139      145      0.0        2 :     p = p * p;
   401      541      0.0        0 :     p = p + v;
                                 :     io.output("r", p, dfeFloat(8,24));
                                 :   }
                                 : }
```

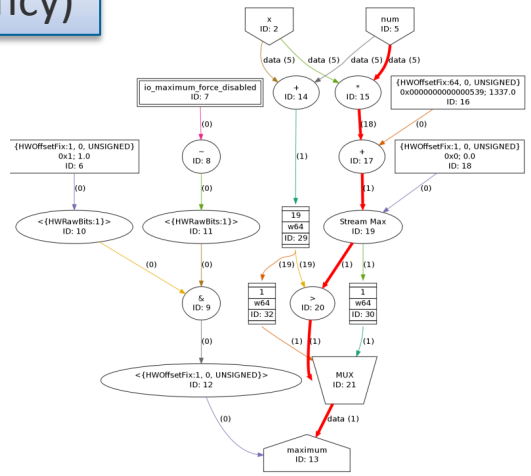
Timing annotation

- MaxCompiler gives detailed latency and area annotation back to the programmer inside MaxIDE

```
27          :
28      12.8:       d.Buy = ask.Price <= lowPrice & order_book.securityId === secId;
29          :       d.Sell = bid.Price >= highPrice & order_book.securityId === secId;
30       6.4:       d.Quantity = d.Buy ? ask.Quantity : bid.Quantity;
31          :       d.Price = d.Buy ? ask.Price : bid.Price;
32          :
```

12.8ns + 6.4ns = 19.2ns (total compute latency)

- Evaluate precise effect of code on latency and chip area



MAXELER Technologies

Measuring Utilization

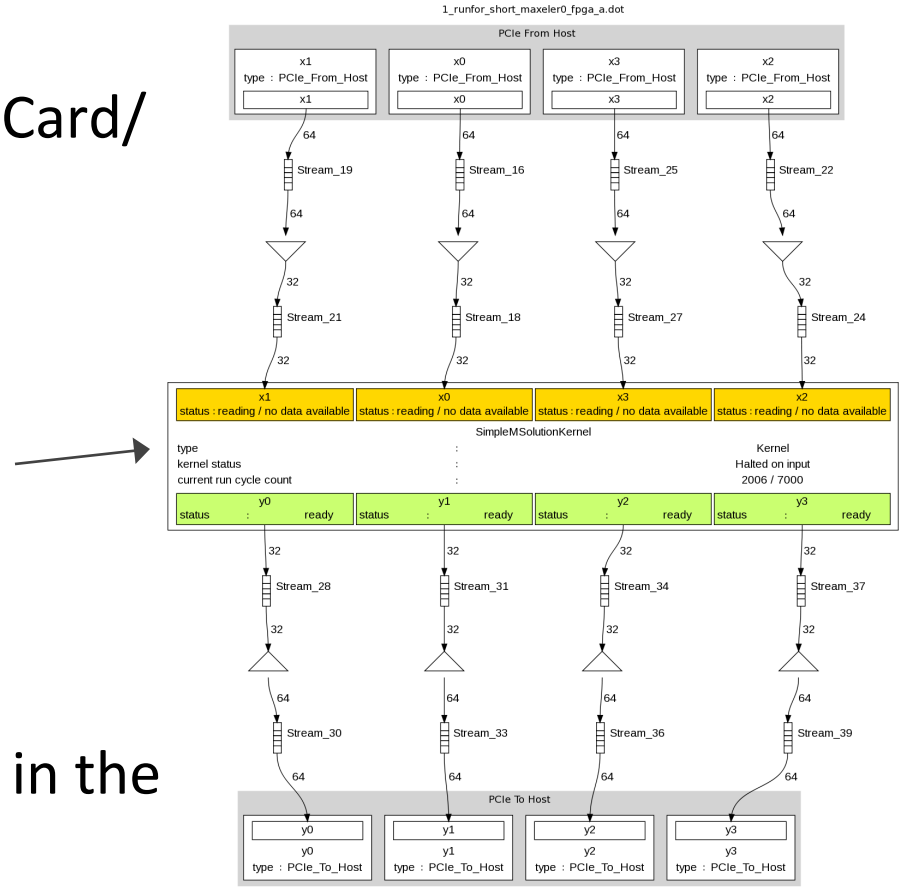
- *Top* measures % of time CPU is running
- *Maxtop* monitors % of time the DFE is running

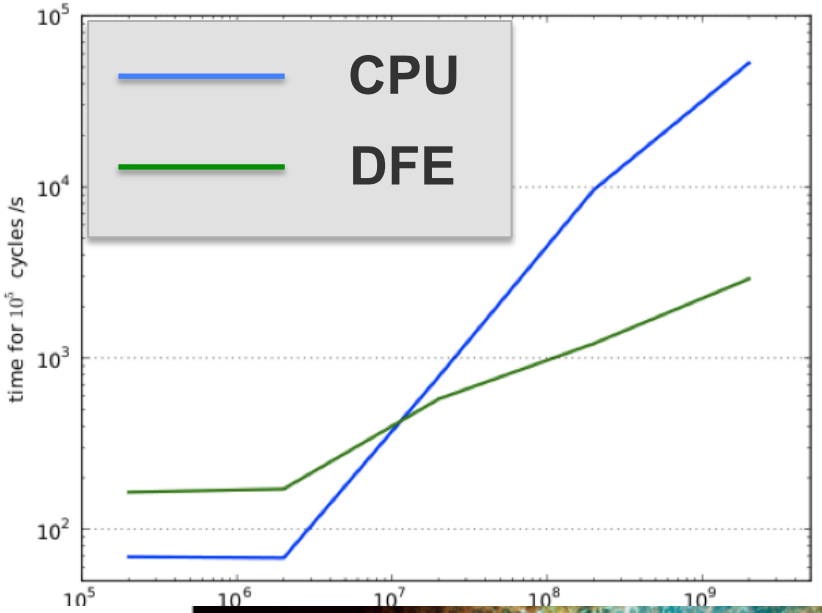
```
MaxTop Tool 2011.2
Found 2 Maxeler card(s) running MaxelerOS 2011.2
Card 0: MAX3A (P/N: 13424) S/N: 219270088 Mem: 24GB DFE(s): 1 /dev/maxeler0
Card 1: MAX3A (P/N: 13424) S/N: 000025559 Mem: 24GB DFE(s): 1 /dev/maxeler1

DEVICE      %DFE    TEMP   BITSTREAM      PID    USER        TIME       COMMAND
maxeler0    66.6%   57.1C  9d9de1...      12333  jspooner    00:00:39   model
maxeler1    0.0%    54.6C  9d9de1...      -      -           -          -
```

Debugging

- MaxCompiler Simulation
  - Simulate a complete MaxCard/ Host system
  - Cycle-accurate
  - Bit-accurate
  - ~100x faster than HDL simulation
- MaxDebug Hardware Debugging
  - See the status of streams in the DFE

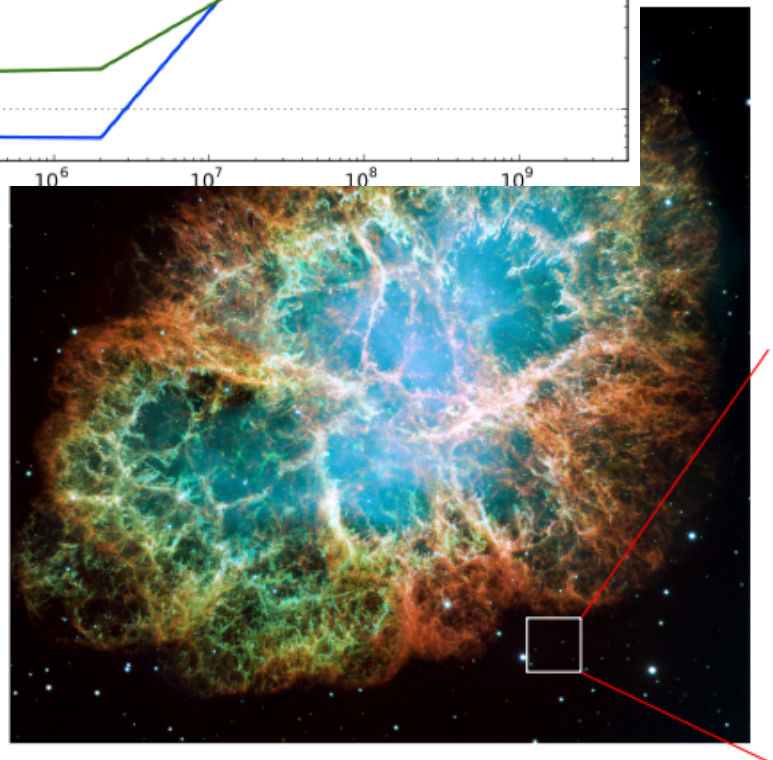


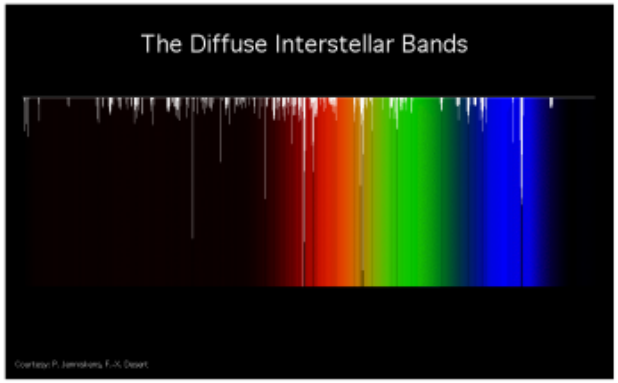


UNIVERSITY OF CAMBRIDGE

Imperial College London

Prof Alex Thom, use of DFEs for Astrochemistry?





http://www.nasa.gov/

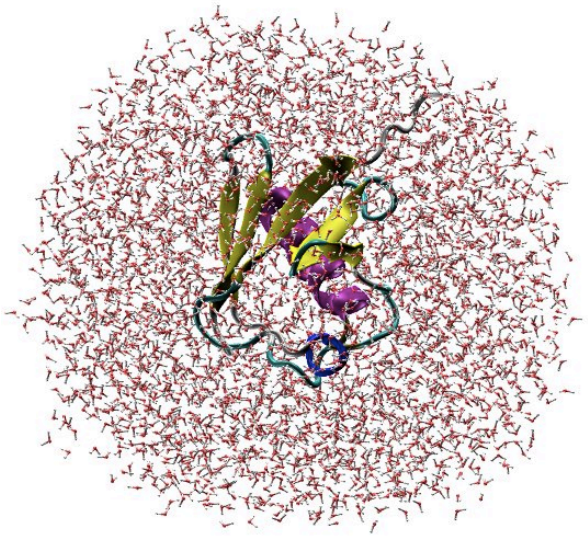
P. Jenniskens and F.-X. Desert, *Astronomy and Astrophysics Supp. Ser.* **106**, 39–78 (1994)

MAXELER Technologies

Molecular Modelling

Computational chemistry allows to:

- determine molecular structures
- relate molecular composition to its function
- clarify catalytic mechanisms
- design new materials

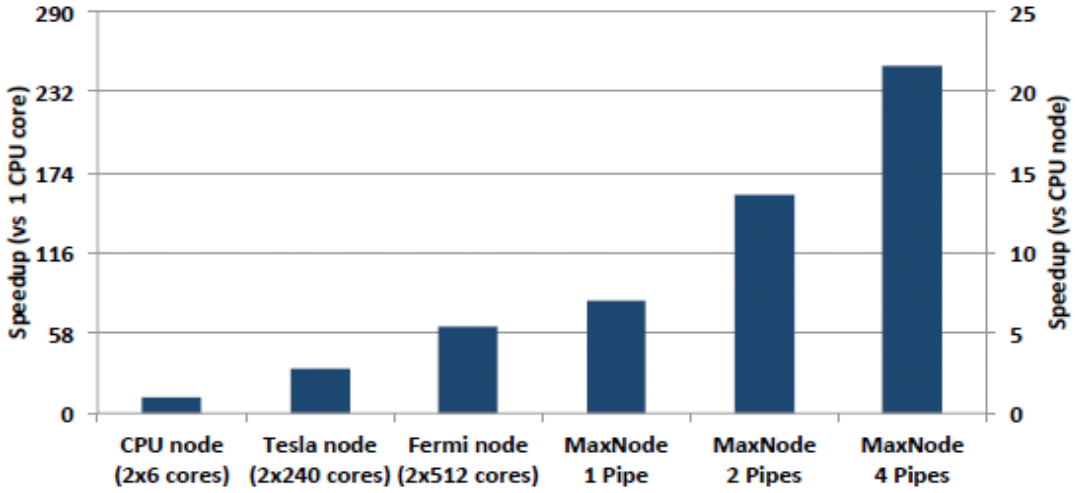


Molecular Mechanics is the method of choice for the computation of large molecular systems (> 1000 atoms).

AMOEBA Model:
Electrostatic induced dipole

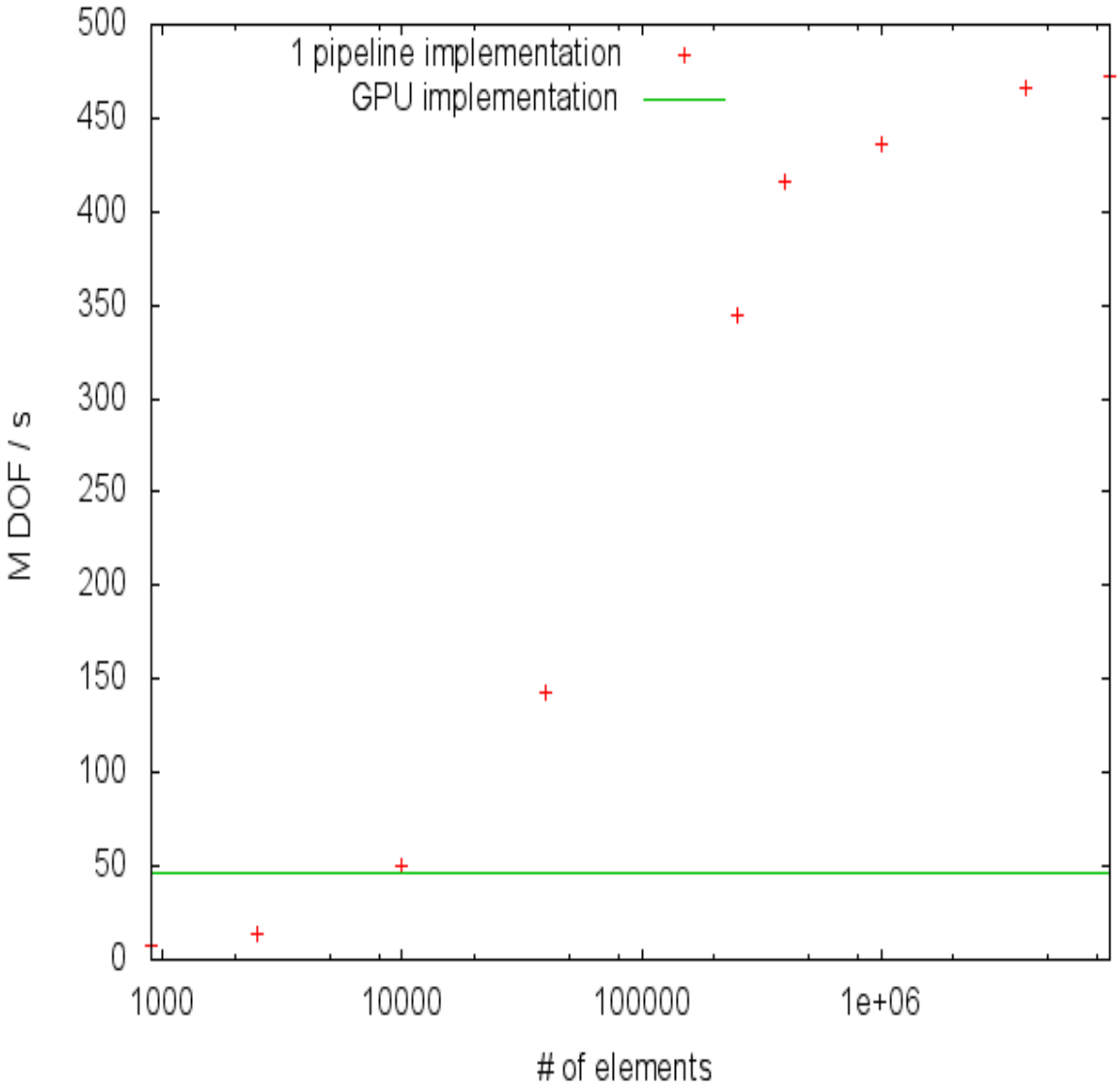
$$\mu_{i,\beta}^{ind} = \alpha_i \left(\sum_j \mathbf{\Gamma}_\beta^{ij} M_j + \sum_{k,\gamma} T_{\beta\gamma}^{ik} \mu_{k,\gamma}^{\text{ind}} \right)$$

Publicly available software
TINKER
http://dasher.wustl.edu/ffe



MAXELER Technologies

CFD Performance



- *Max3A workstation with MAX3 DFE*

- For this 2D linear advection test problem we achieve ca. 450M degree-of-freedom updates per second
- For comparison a GPU implementation (of a Navier-Stokes solver) achieves ca.50M DOFs/s

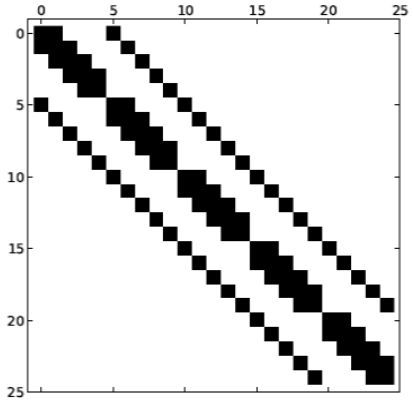
Sparse Matrix Solving

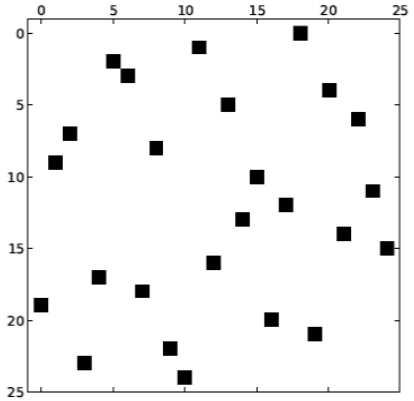
O. Lindtjorn et al, HotChips 2010

- Sparse matrices are used in a variety of important applications
- Matrix solving. Given matrix **A**, vector *b*, find vector *x* in:

$$\mathbf{A}x = b$$

- Direct or iterative solver
- Structured vs. unstructured matrices



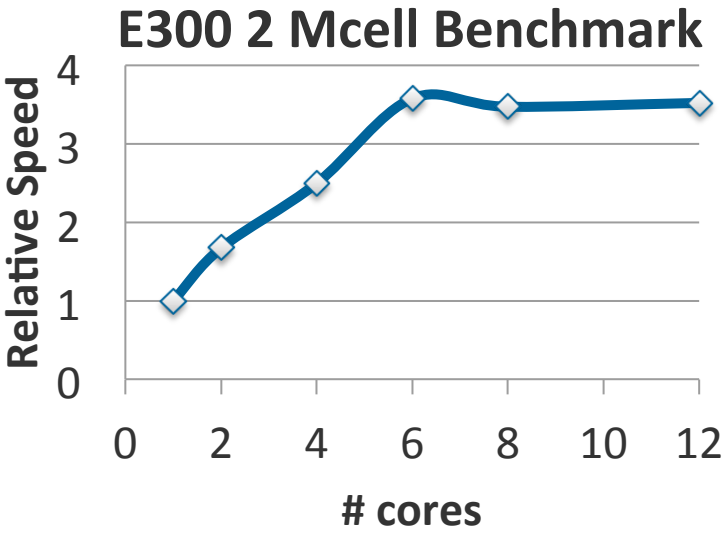


Schlumberger

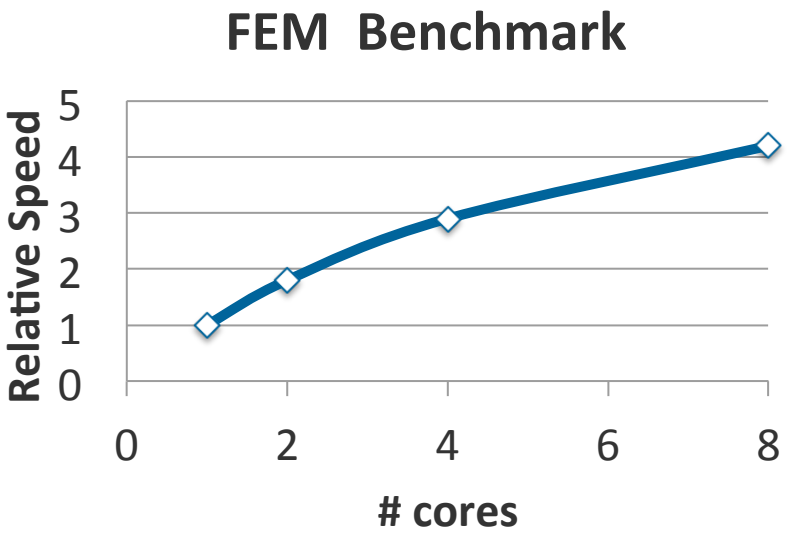
MAXELER Technologies

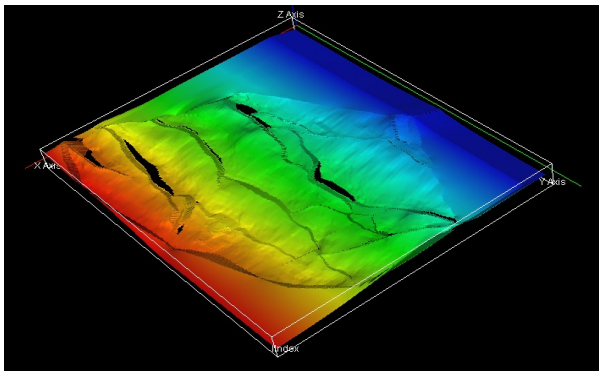
Typical Scalability of Sparse Matrix

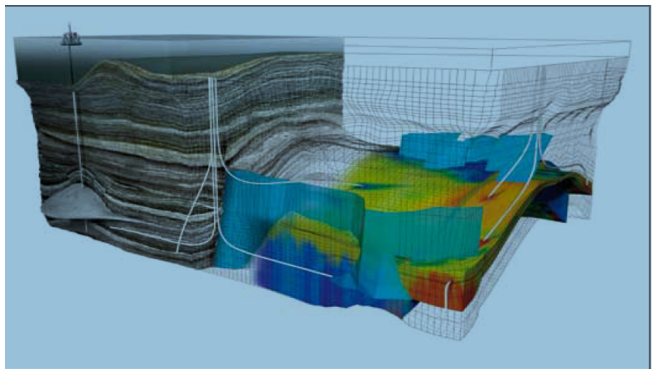
Eclipse Benchmark
(2 node Westmere 3.06 GHz)



Visage – Geomechanics
(2 node Nehalem 2.93 GHz)



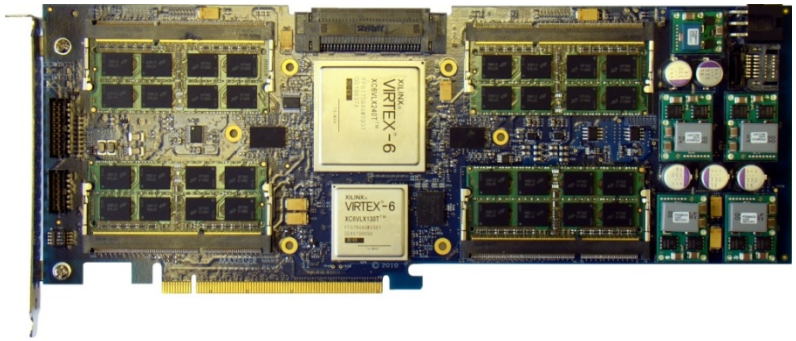


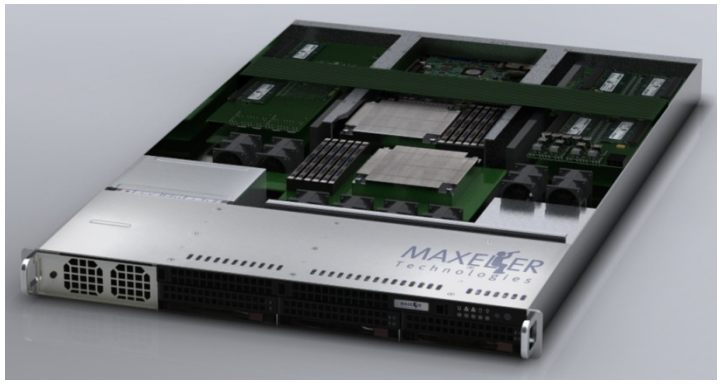


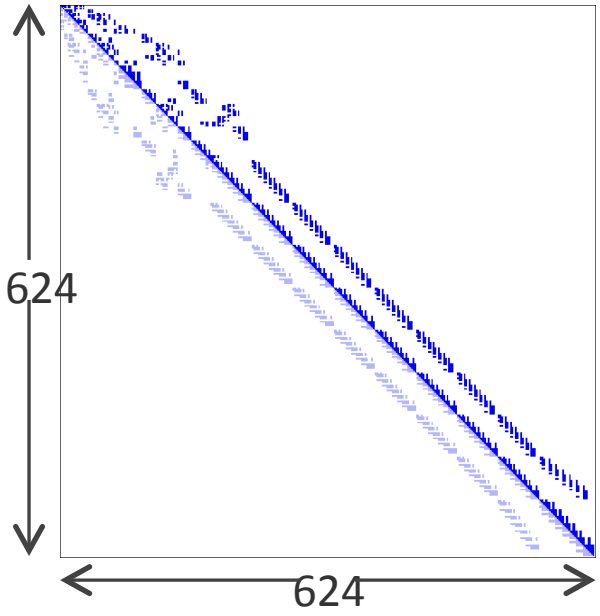
Schlumberger

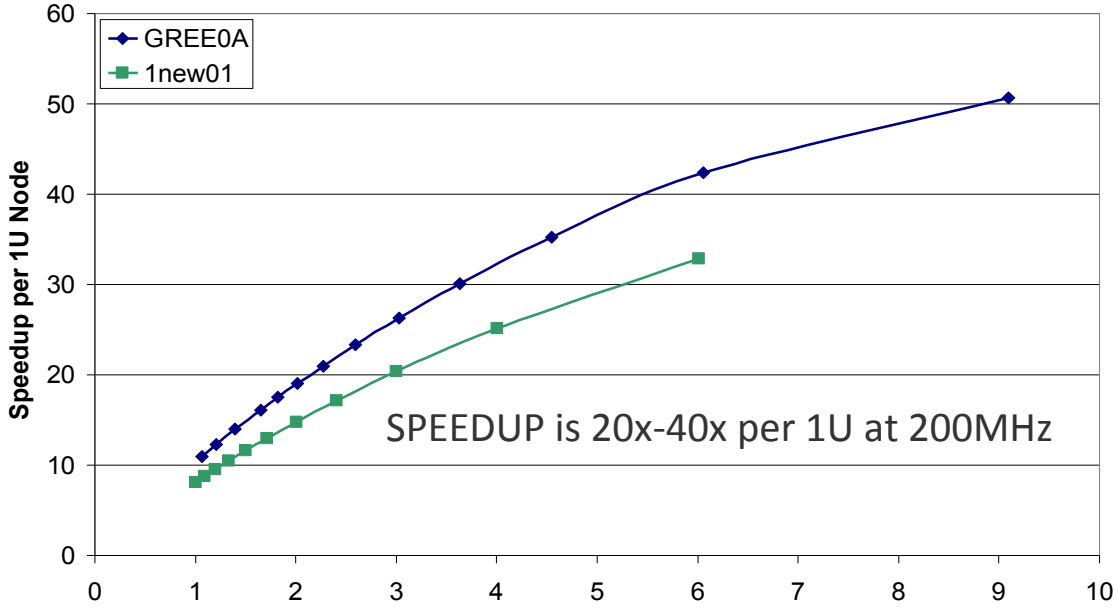
MAXELER Technologies

Sparse Matrix on DFEs



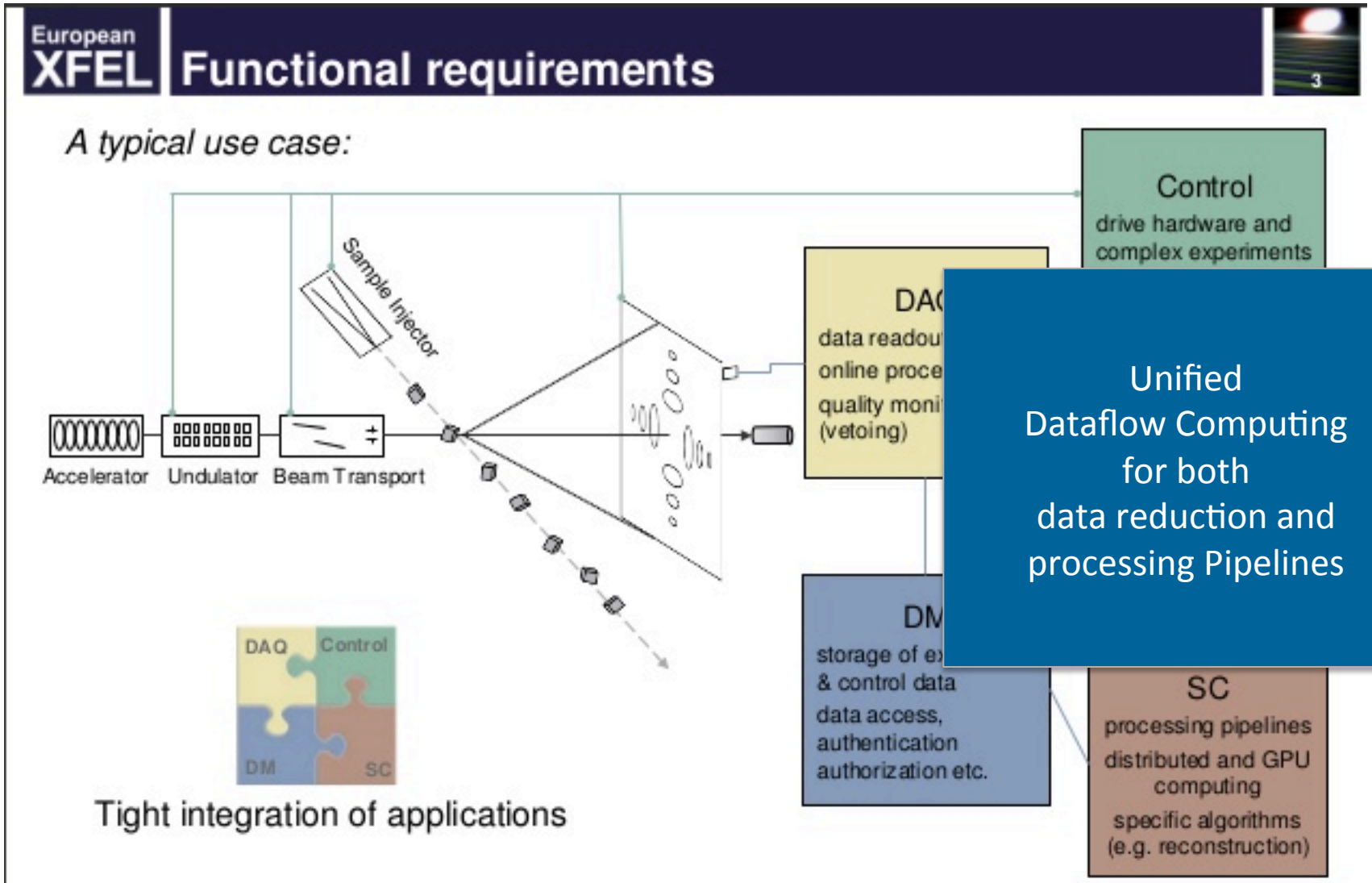






Maxeler Domain Specific Address and Data Encoding

Computing at the Sensor and Beyond



Why use Maxeler DFEs for Physics Experiments

1. Maxeler Dataflow computers are similar to Physics experiment
2. Funded by JP Morgan, Schlumberger, Chevron, CME Group, …
3. 5th generation (MAX5) commercial offering coming up next year
4. Stable and production ready infrastructure
5. All IP is protected by Iron Mountain archiving, used as part of the backbone of the financial system

6. Maximize re-use, and reduce risk

150 Maxeler University Program Members