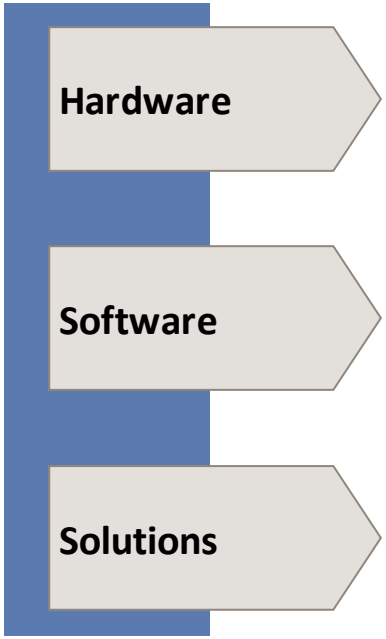
COMPUTING WITH DATAFLOW ENGINES

Robin Bruce
November 2012

About Maxeler Technologies

- Maxeler offers complete solutions for high performance computing problems
- Founded 2003, funded by our clients
- ~70 people, offices in London, UK and Palo Alto, CA



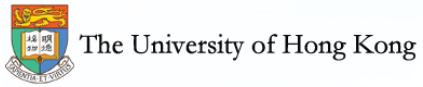
- Dataflow Engines (DFE): Reconfigurable chip, lots of memory
- Dataflow Node: 1U solutions with multiple DFEs
- Rack: 10U, 20U or 40U, balancing compute, storage & network

- MaxCompiler: providing Dataflow programmability
- MaxelerOS: Resource management of Dataflow Computing
- Runtime support: memory management and data choreography

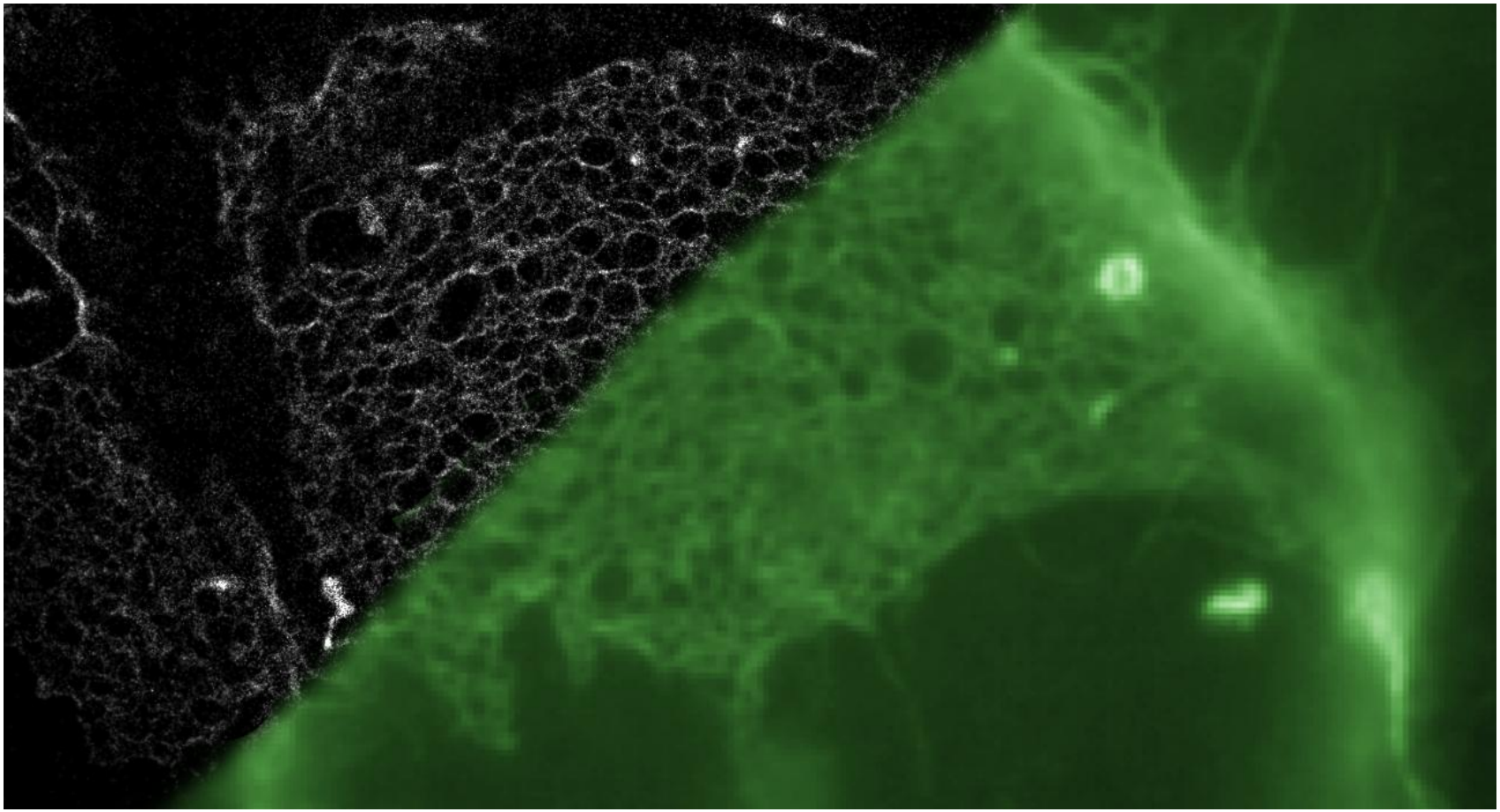
- Analysis and re-architecting of existing applications
- Algorithms and numerical optimization
- Integration into business and technical processes

MAXELER Technologies

Academic Collaboration



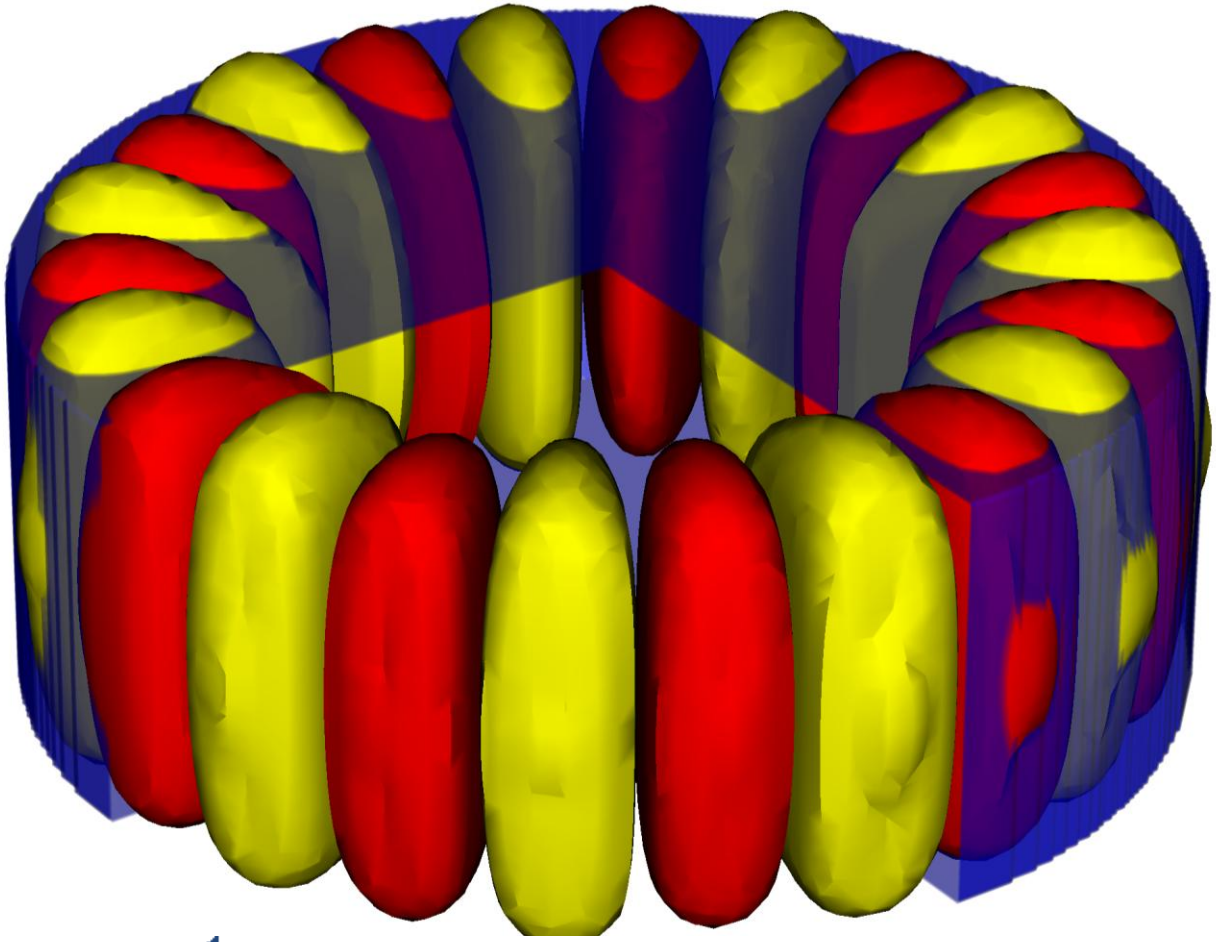
MAXELER
Technologies



Localization microscopy enhances the resolution of fluorescence light microscopy (shown in green) by about an order of magnitude. Single fluorescent molecules act as switchable markers. Their detected signals can be fitted with a two-dimensional Gaussian distribution and thus located with sub-pixel resolution. Using MaxCompiler we achieved a hardware acceleration of 225 in signal detection and fitting.

KIRCHHOFF-INSTITUTE FOR PHYSICS

MAXELER Technologies

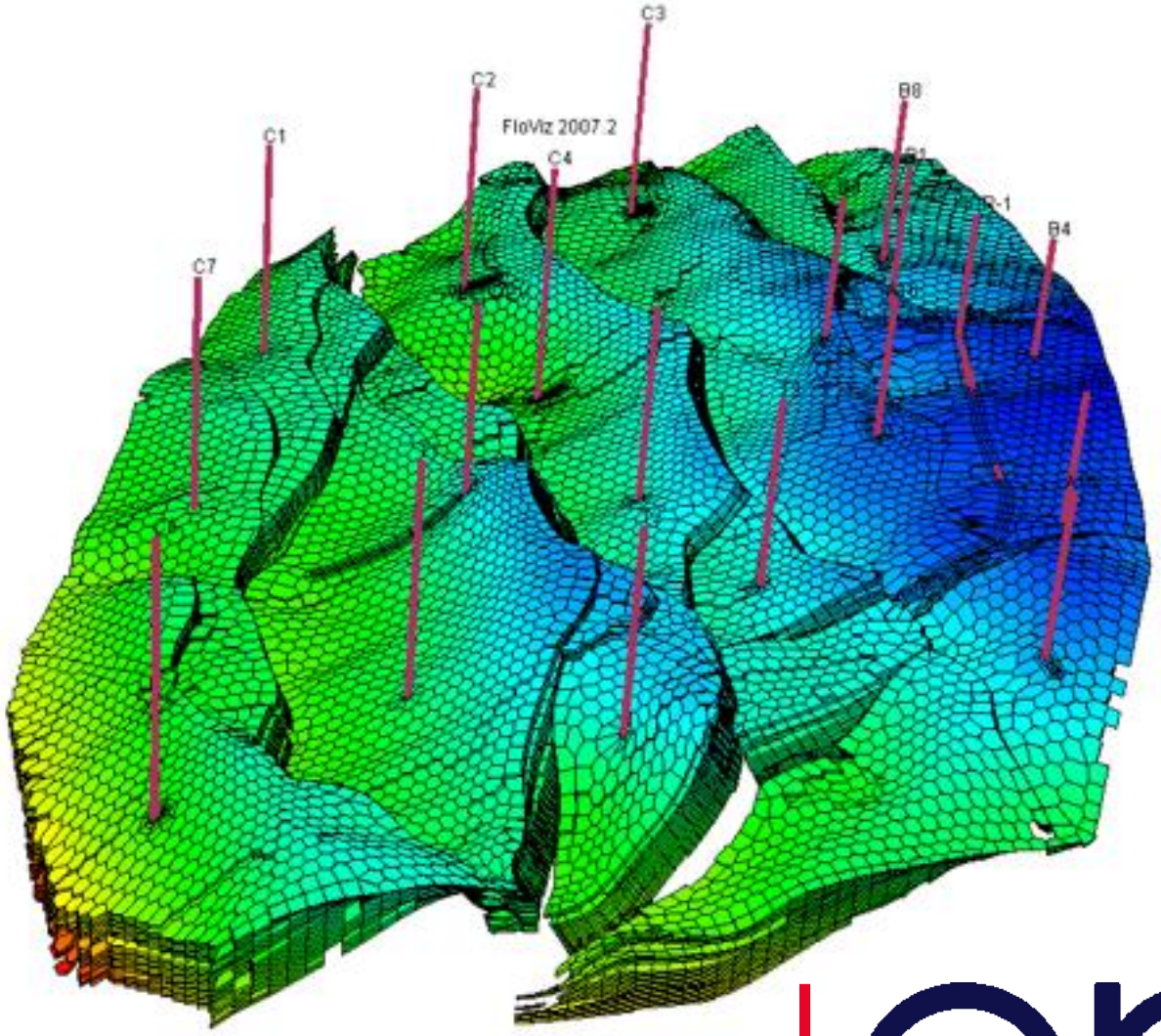


UNIVERSITÄT PADERBORN

Die Universität der Informationsgesellschaft

Acceleration of simulations of computational nanophotonics using Maxeler MaxGenFD Finite Difference Compiler for dataflow

MAXELER Technologies



Hydrocarbon reservoir simulation on Maxeler dataflow computers at the Edinburgh Parallel Computing Centre

THE UNIVERSITY OF EDINBURGH

epcc

MAXELER Technologies

CAVE automatic virtual environment

This project will investigate the use of dataflow architectures in creation of *high-end immersive virtual reality*. The goal is to build a proof of concept simulator with *extremely low latency*.

UCL

MAXELER Technologies

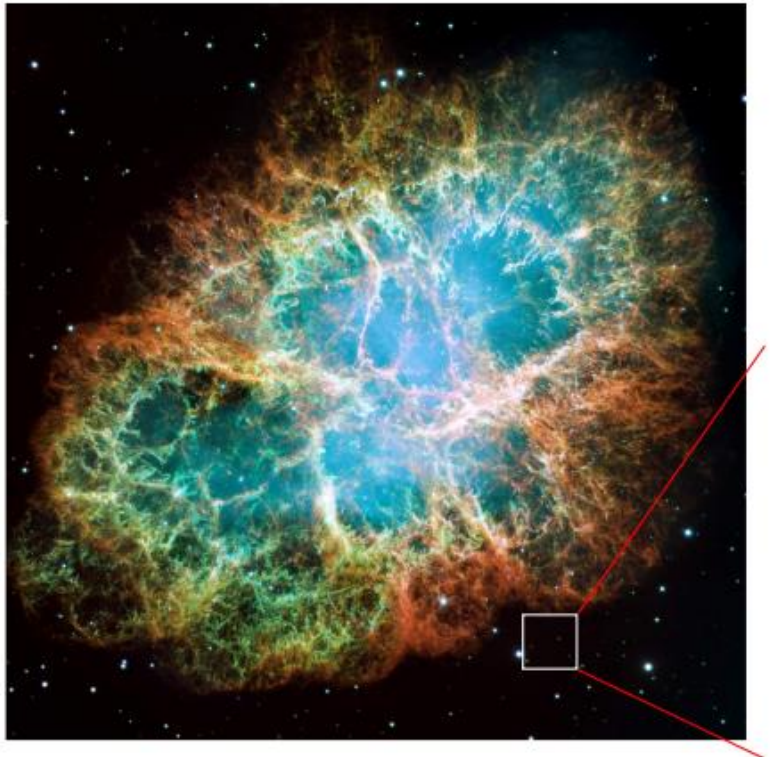
A MAX-UP project example

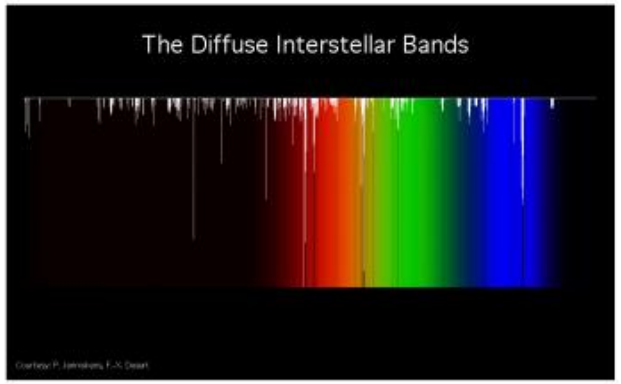
Imperial College London

Astrochemistry

MAXELER Technologies
MAXIMUM PERFORMANCE COMPUTING

coronene: $C_{24}H_{12}$





http://www.nasa.gov/

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

MAXELER Technologies

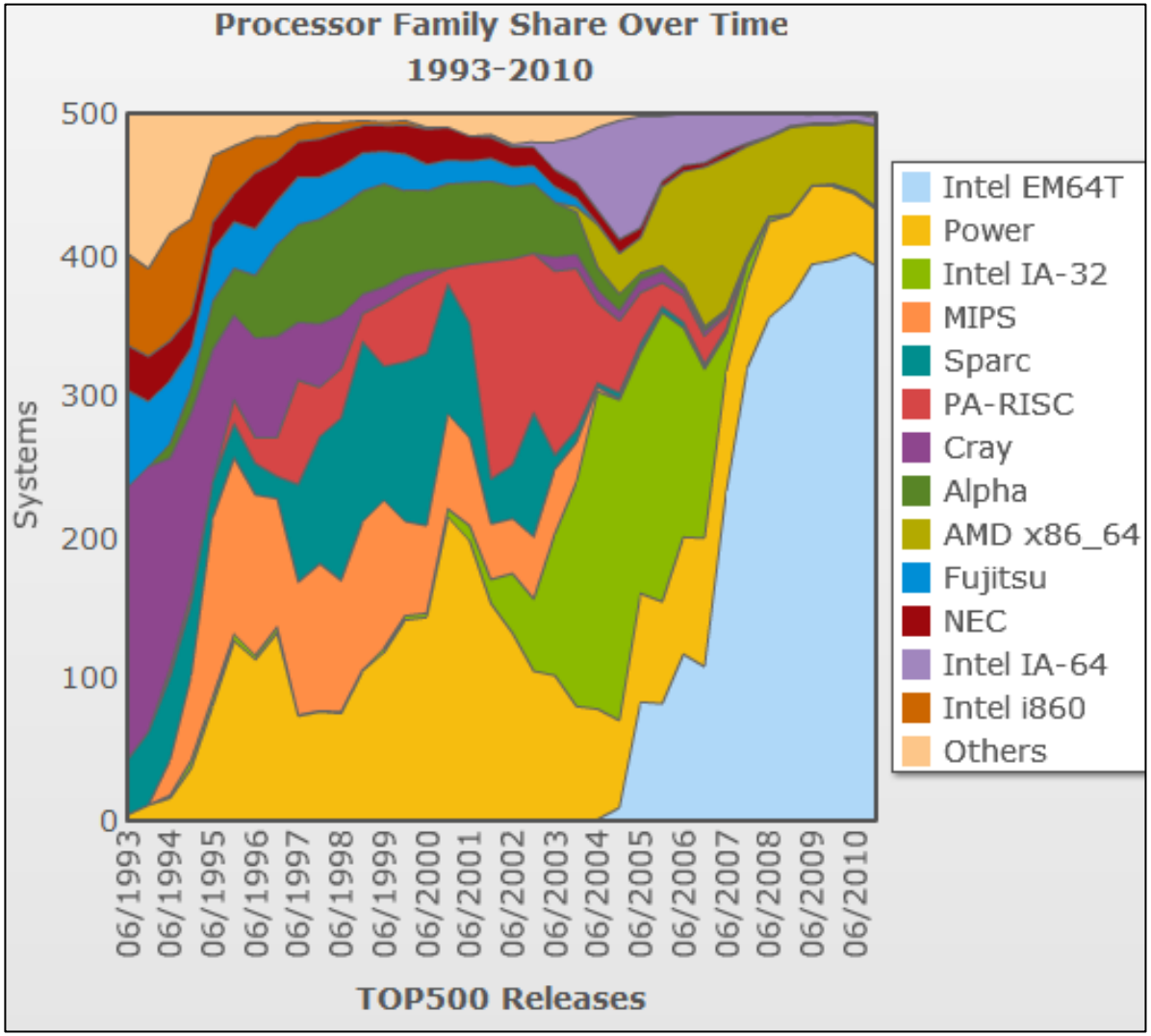
Overview

Reconfigurable dataflow computing

Solving the programming challenge

Real world impact

Rise of x86 Supercomputers



MAXELER Technologies

The Exaflop Supercomputer (2018)

- 1 exaflop = 10
- Using processo t 2.5GHz
- **50M CPU cores**
- What about power?
  - Assume power envelope
  - Moore's Law scaling: 6 c
  - 500k CPU chips
- **50MW** (*just for CPUs!*) → **100MW likely**
- 'Jaguar' power consumption: 6MW

How do we program this?

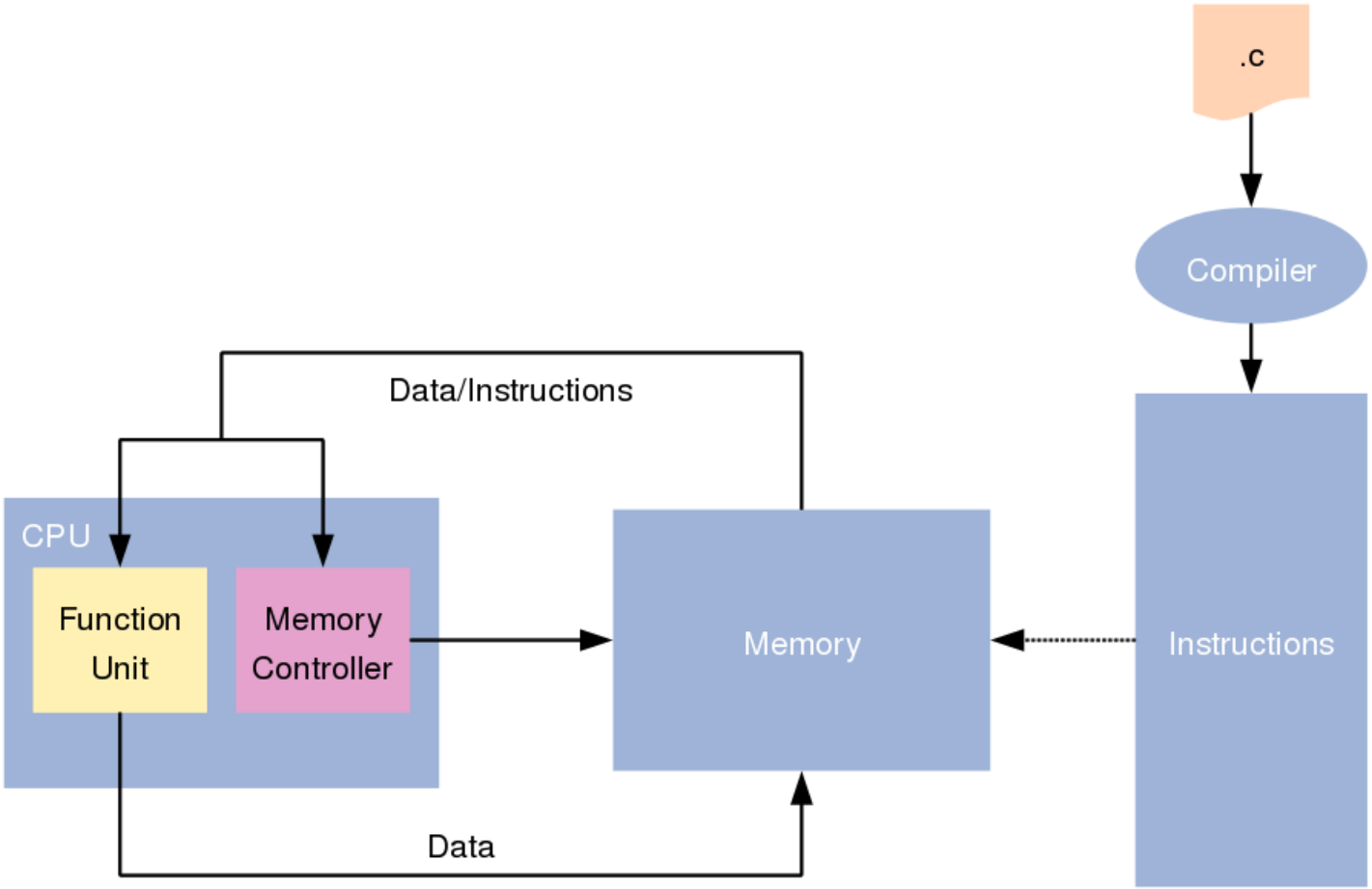
Who pays for this?

Solving Exascale Problems

- Peak performance doesn't matter
- (Whisper it) LINPACK doesn't matter
- The TOP500 doesn't matter
- We're interested in solving *real* problems
  - Science
  - Commerce
  - Industry
- Can we build *application-specific computers?*

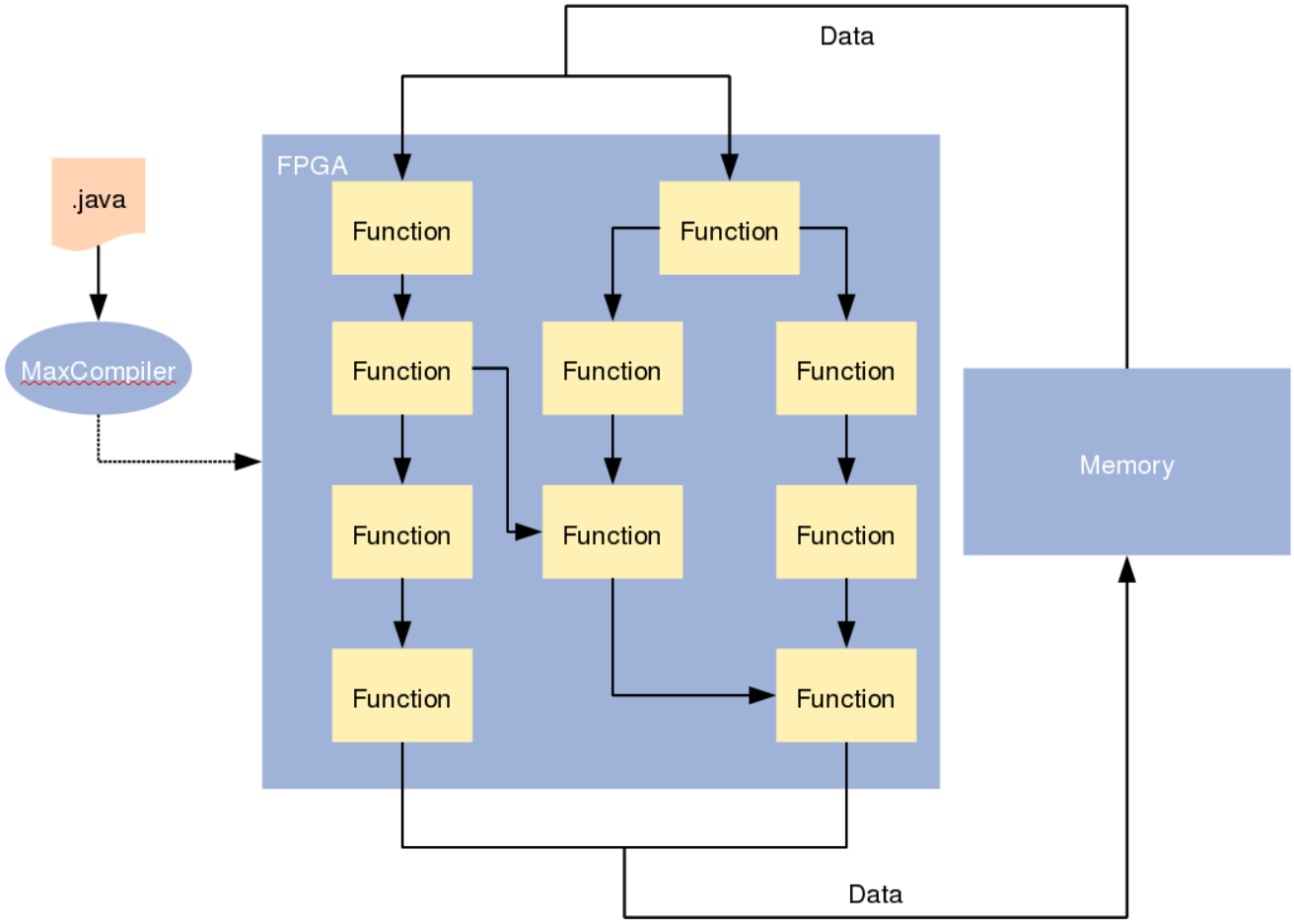
RECONFIGURABLE DATAFLOW COMPUTING

Computing with Instruction Processors



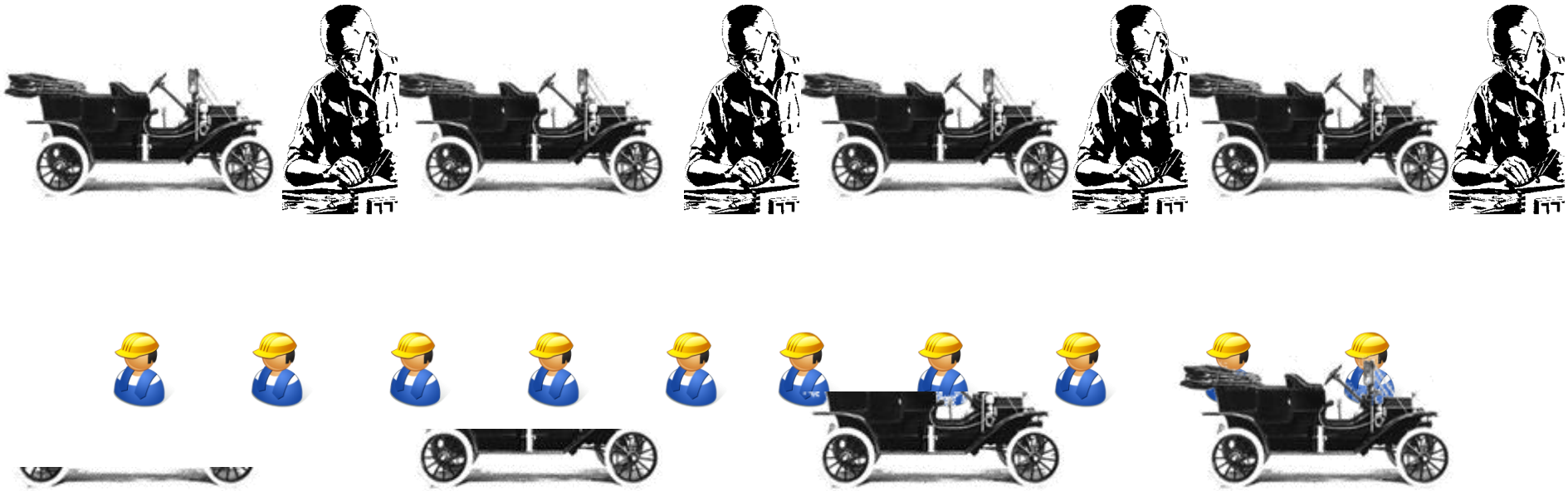
MAXELER Technologies

Computing with Dataflow



MAXELER Technologies

Explaining Control Flow versus Data Flow



- Experts are expensive and slow (control flow)
- Many specialized workers are more efficient (data flow)

MAXELER Technologies

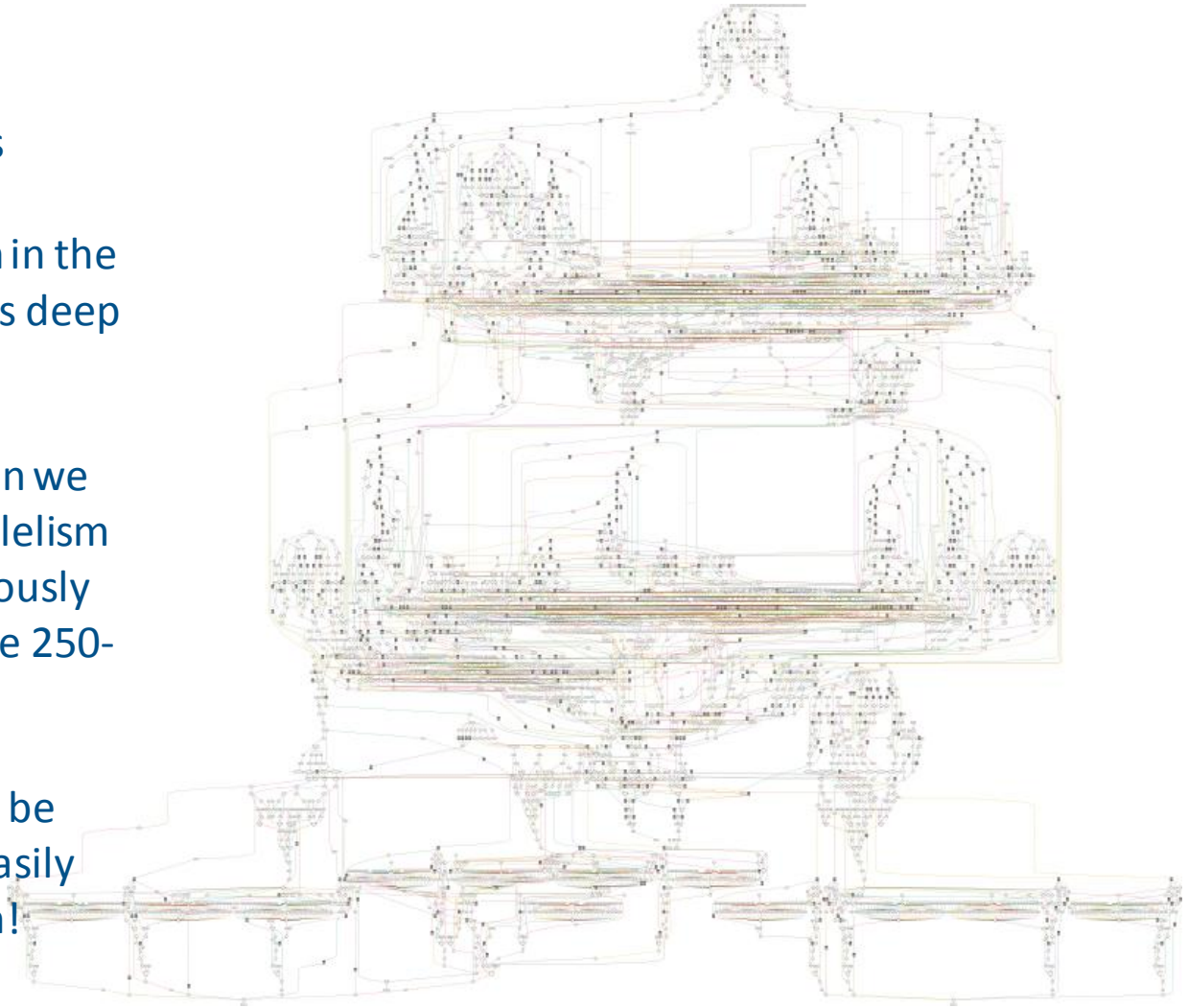
Real-life MaxCompiler Dataflow Graph

Consists of 4866 nodes

We have pipeline parallelism in the vertical dimension. 250 cycles deep on the critical path.

On the horizontal dimension we have additional spatial parallelism up to 100 nodes simultaneously processing at one point in the 250-level pipeline.

Obviously too complex to be generated manually and easily understood by a human!



MAXELER Technologies

Maxeler Hardware Solutions 2012

MPC C Series

CPUs plus DFEs

Intel Xeon CPU cores and up to 6 DFEs with 288GB of RAM

MPC X Series

DFEs shared over Infiniband

Up to 8 DFEs with 384GB of RAM and dynamic allocation of DFEs to CPU servers

MPC N Series

Low latency connectivity

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

MaxWorkstation

Desktop development system

MaxCloud

On-demand scalable accelerated compute resource, hosted in London

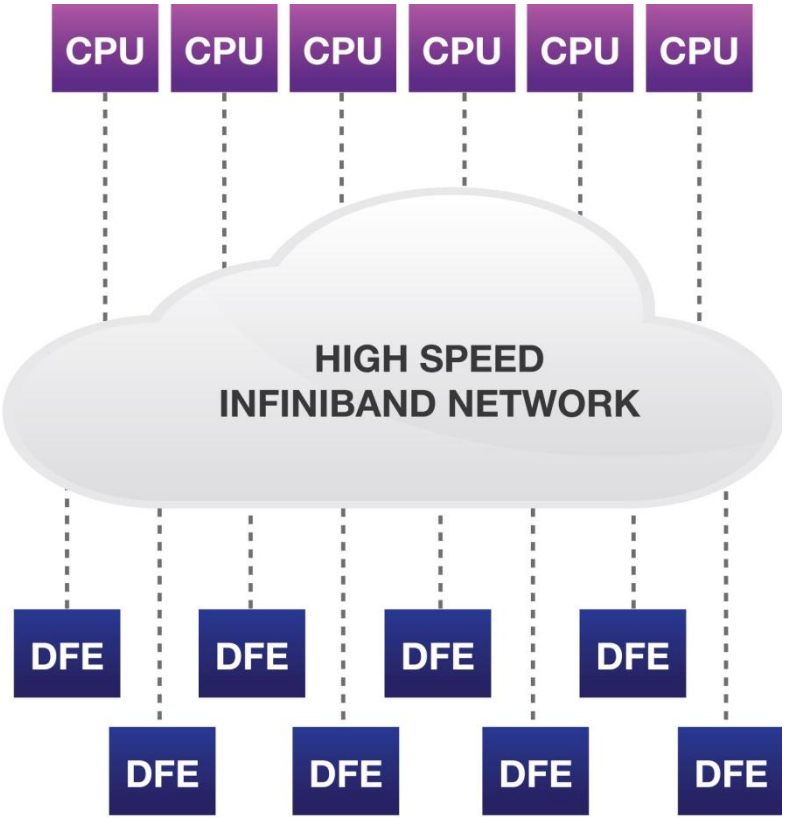
MAXELER Technologies

MPC X Series

1U dataflow cloud providing dynamically scalable compute capability over Infiniband

MPC-X1000

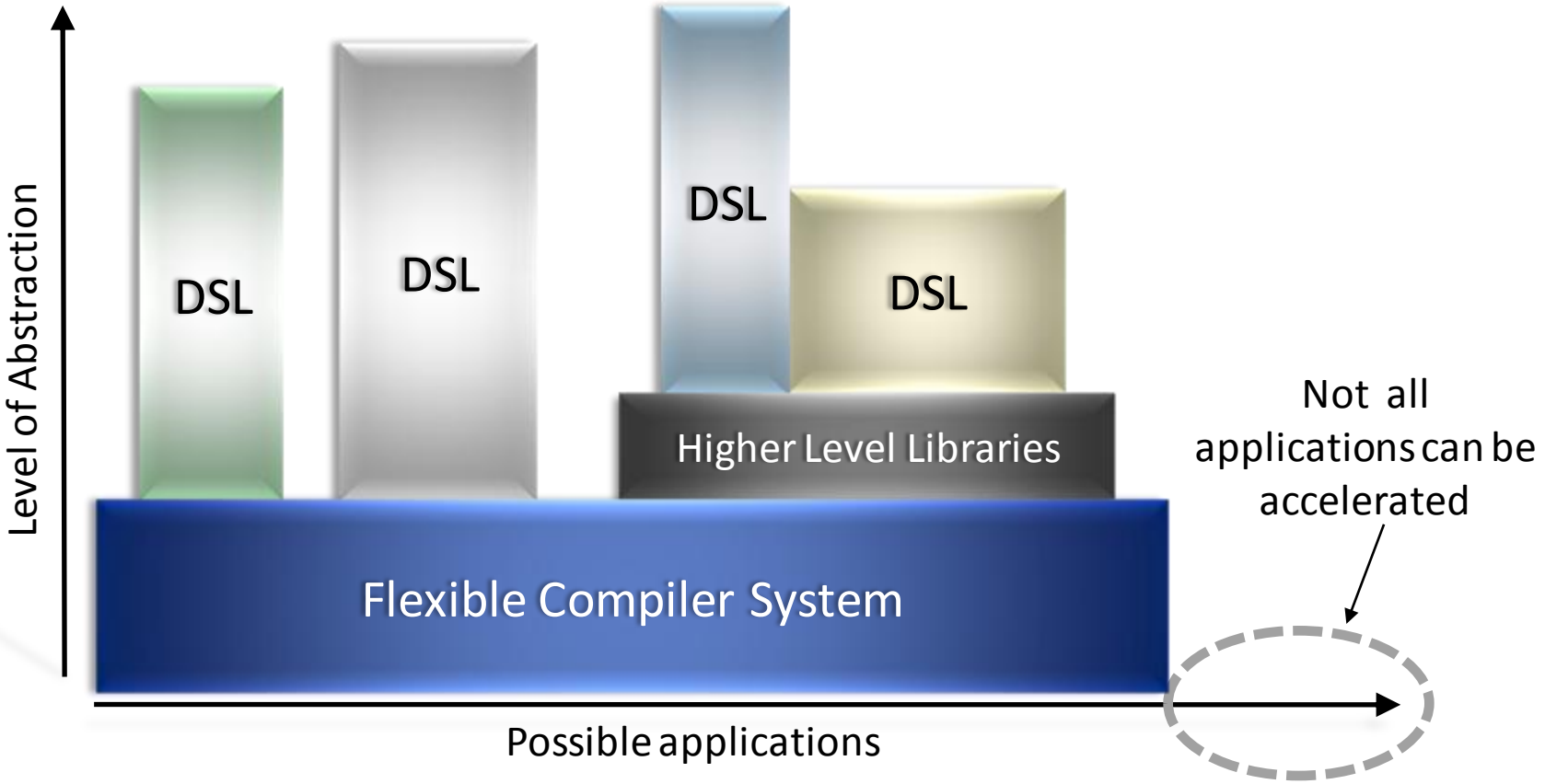
- 8 *vectis* dataflow engines (DFEs)
- 192GB of DFE RAM
- Dynamic allocation of DFEs to conventional CPU servers
  - Zero-copy RDMA between CPUs and DFEs over Infiniband
- Equivalent performance to 40-60 x86 servers



MAXELER Technologies

SOLVING THE PROGRAMMING CHALLENGE

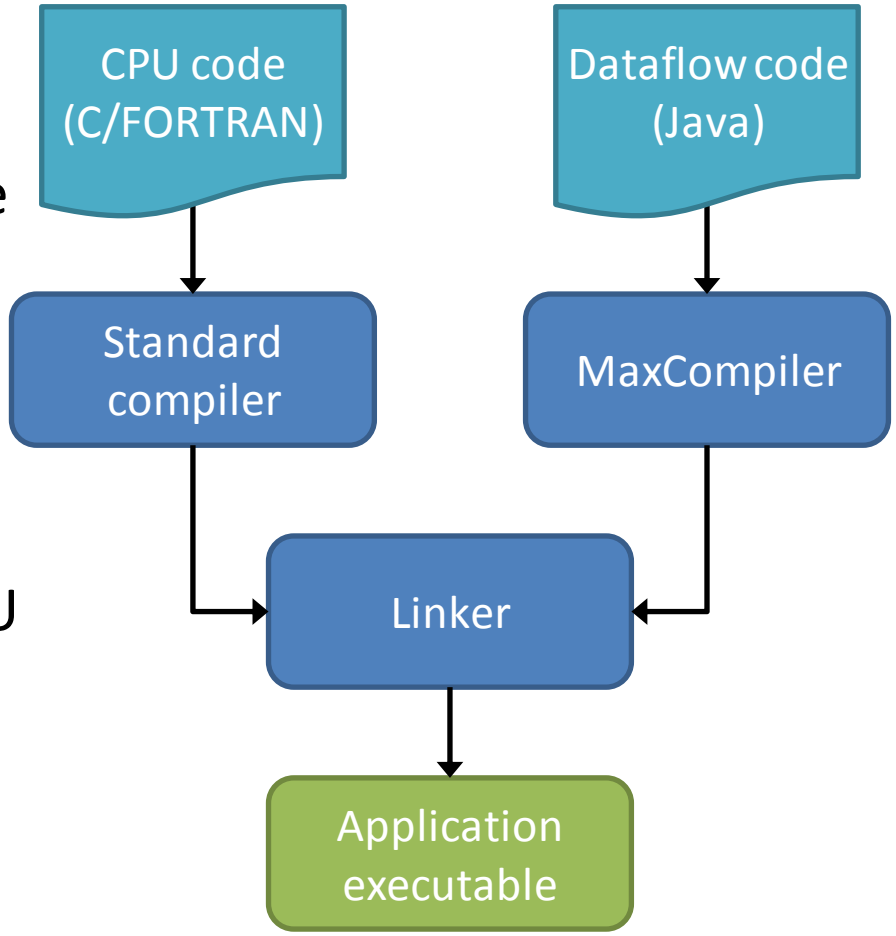
Accelerator Programming Models



MAXELER Technologies

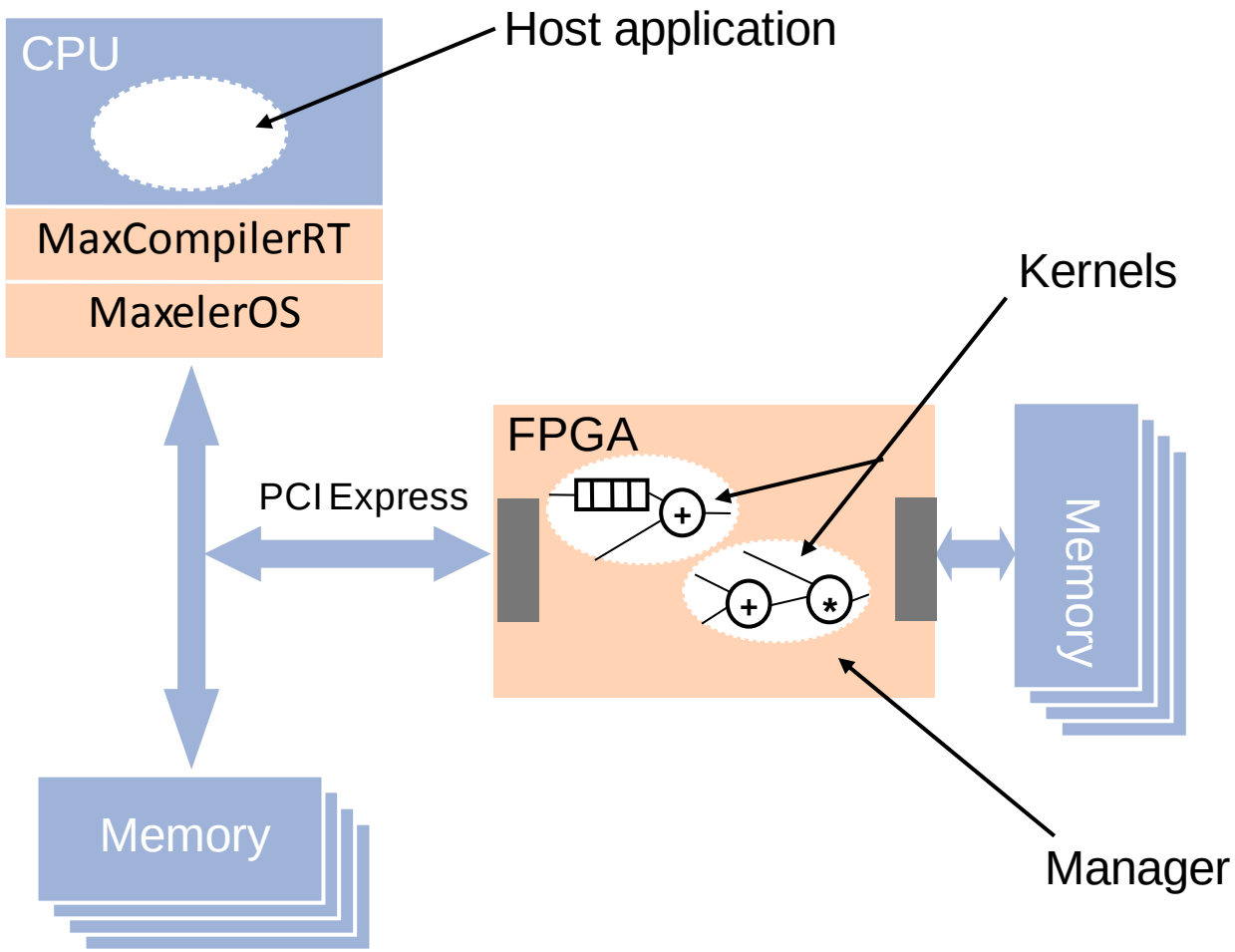
Programming Dataflow Computers

- Programmers:
  - Extract computationally intensive *kernels* and rewrite them as dataflow code
  - Modify CPU code to call the dataflow engine function
- Dataflow code compiles with *MaxCompiler* and links with CPU code to create accelerated application executable
- Application runs normally but with dataflow acceleration



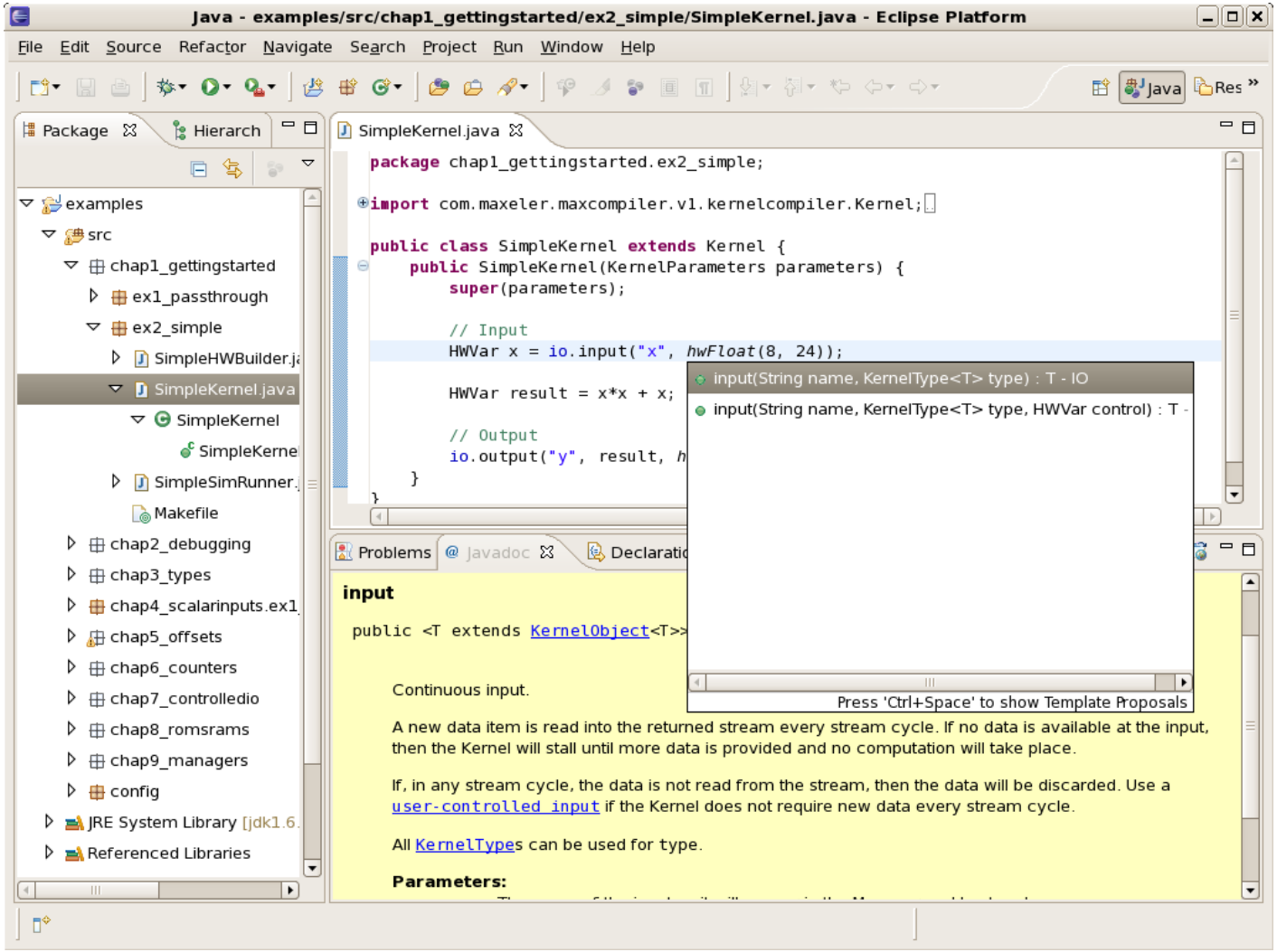
MAXELER Technologies

MaxCompiler: Runtime View



MAXELER Technologies

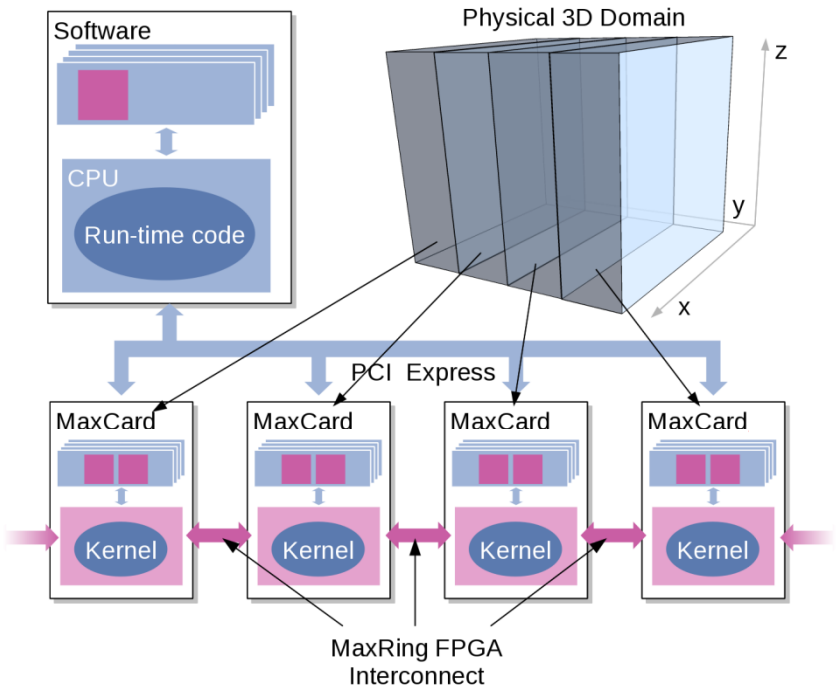
MaxIDE – Dataflow Development



MAXELER Technologies

3D Finite Difference, MaxGenFD

- High level library for quickly and easily accelerating 3D finite difference applications
  - Succinct program code
  - Handles large data sets
  - Automatic domain decomposition and parallelization
  - Stencil optimizations



MAXELER Technologies

Example MaxGenFD, Easy to Code

Host Code (.c)

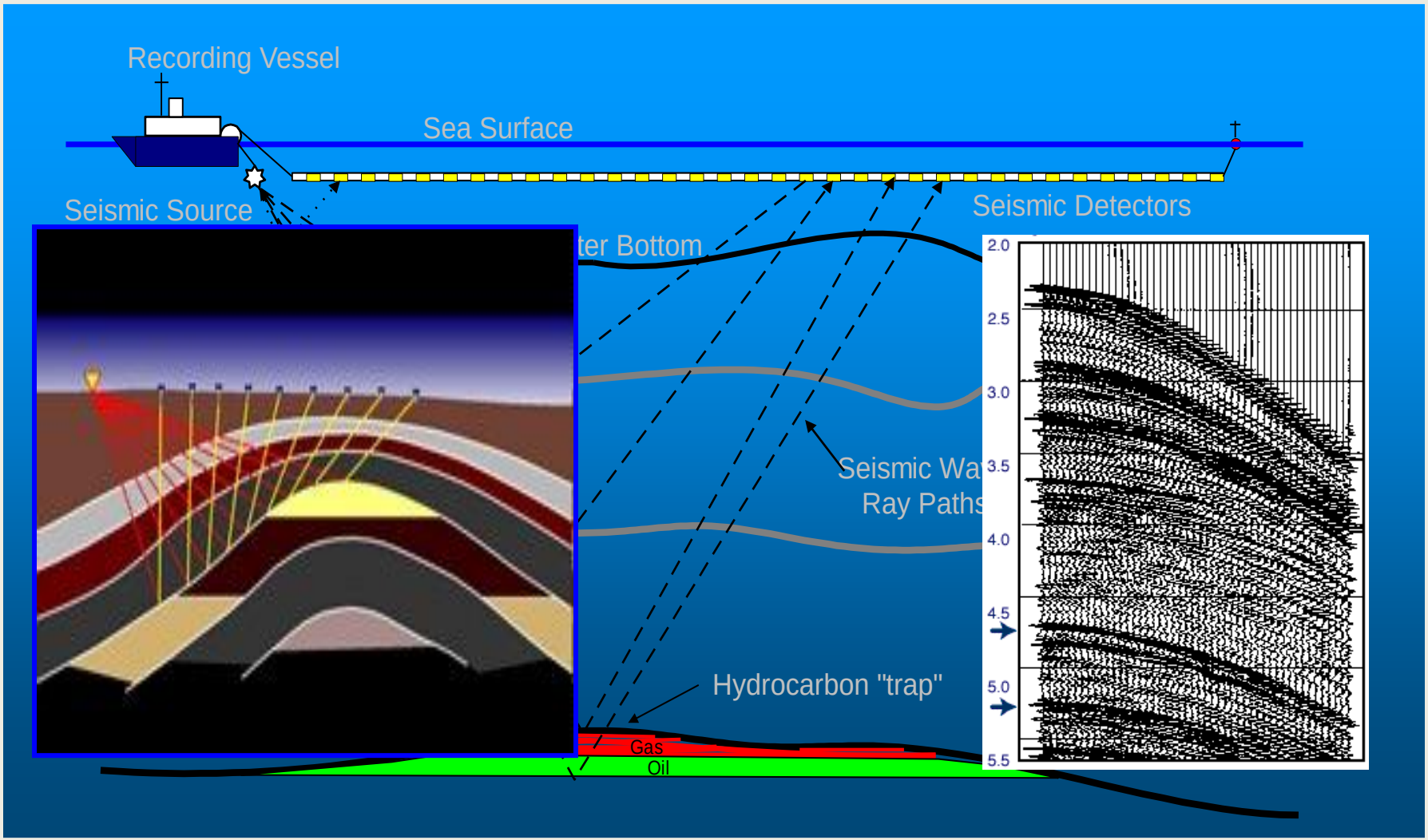
```
for(t = 1; t < tmax; t++) {
    // Set-up timestep
    if (t < tsrc) {
        source = generate_source_wavelet(t);
        maxlib_stream_region_from_host(maxlib, "source", source
            srcx, srcy, srcz, srcx+1, srcy+1, srcz+1);
    }
    maxlib_stream_from_dram(maxlib, "curr", curr_ptr);
    maxlib_stream_from_dram(maxlib, "prev", prev_ptr);
    maxlib_stream_earthmodel_from_dram(maxlib, dvv_array);

    maxlib_stream_to_dram(maxlib, "next", next_ptr);

    maxlib_run(maxlib); // Execute timestep

    swap_buffers(prev_ptr, curr_ptr, next_ptr);
}
```

FDKernel (.java)

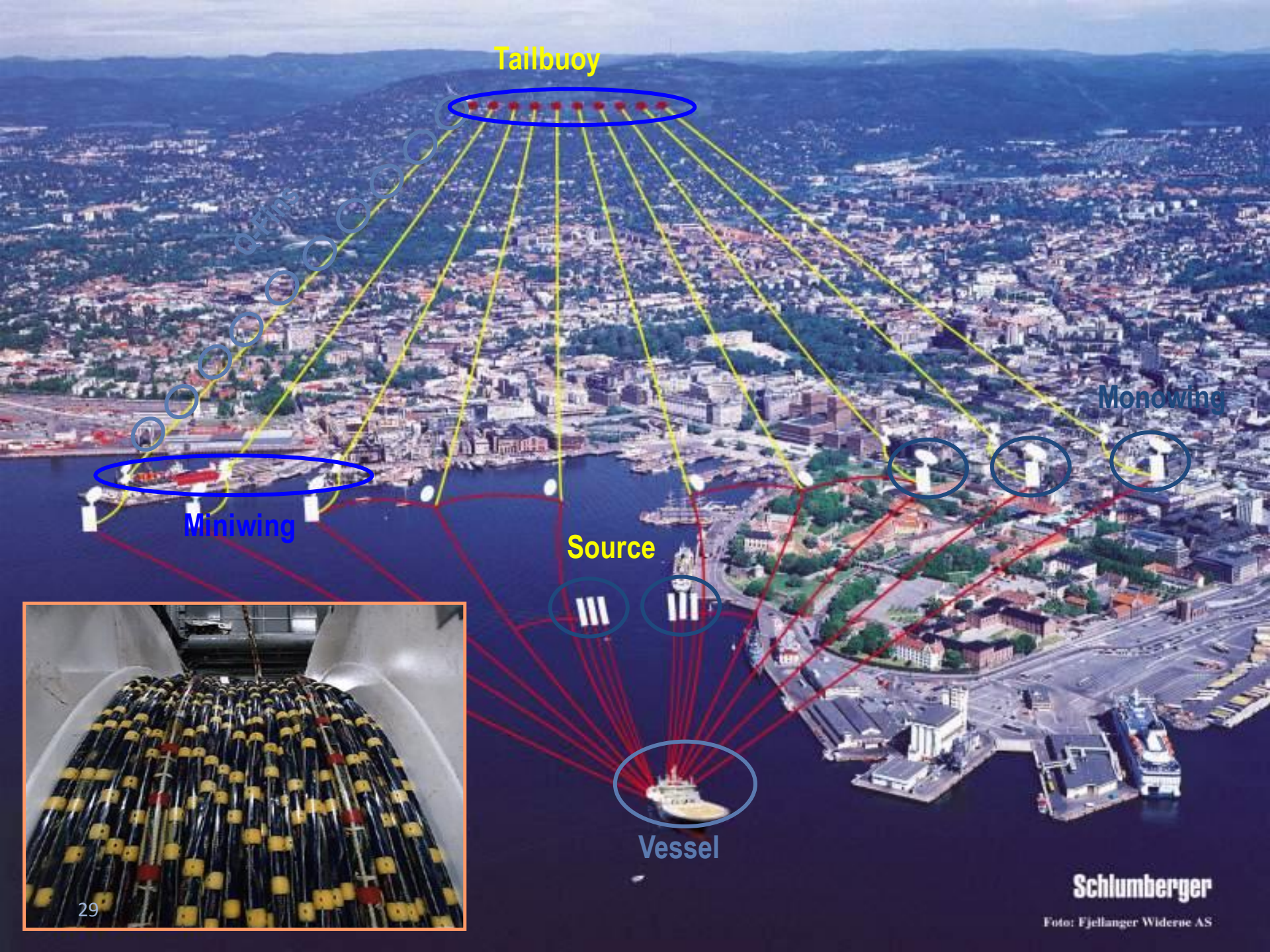
```
public class IsotropicModelingKernel extends FDKernel {
    public IsotropicModelingKernel(FDKernelParameters p) {
        super(p);
        Stencil stencil = fixedStencil(-6, 6, coeffs, 1/8.0);
        FDVar curr = io.wavefieldInput("curr", 1.0, 6);
        FDVar prev = io.wavefieldInput("prev", 1.0, 0);
        FDVar dvv = io.earthModelInput("dvv", 9.0, 0);
        FDVar source = io.hostInput("source", 1.0, 0);

        FDVar l = convolve(curr, ConvolveAxes.XYZ, stencil);

        FDVar next = curr * 2 - prev + dvv * l + source;


        io.wavefieldOutput("next", next);
    }
}
```

REAL WORLD IMPACT

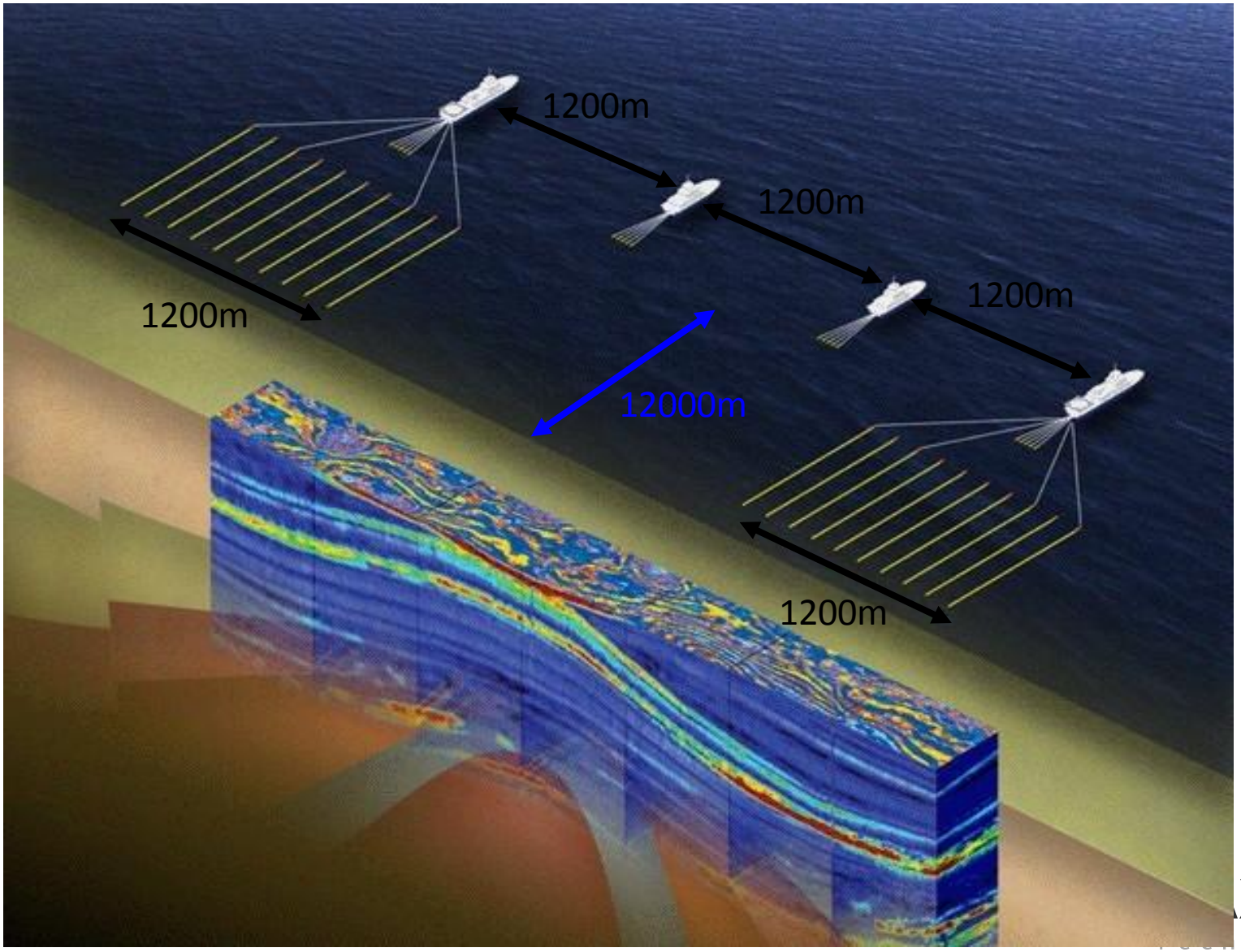
Oil and Gas Exploration



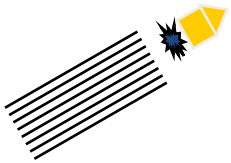
MAXELER Technologies

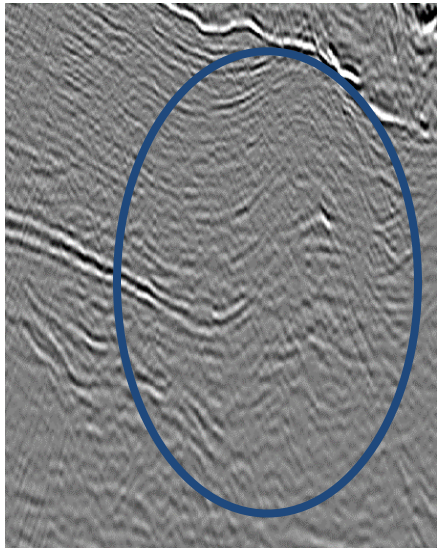

Tailbuoy
Miniwing
Source
Monowing
Vessel
Schlumberger
Foto: Fjellanger Widerøe AS

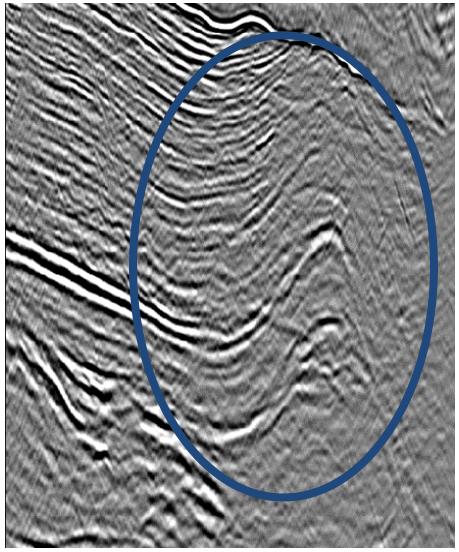
Wide Azimuth Acquisition

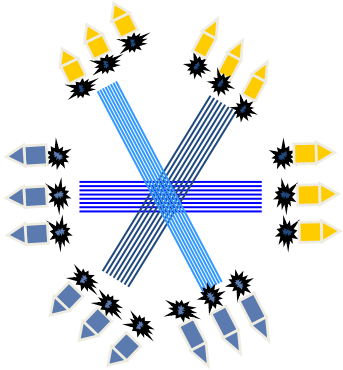


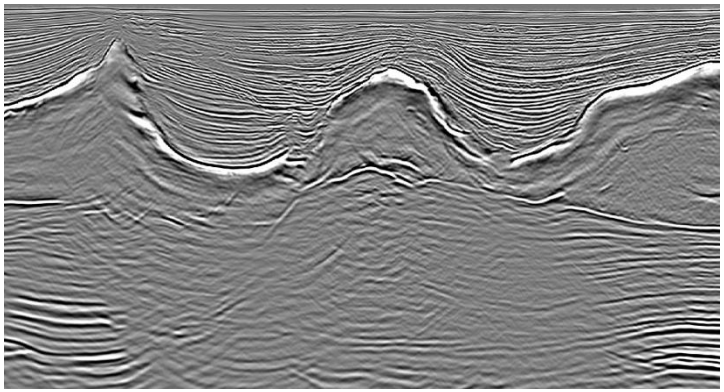
Data Intensity and Complex Physics



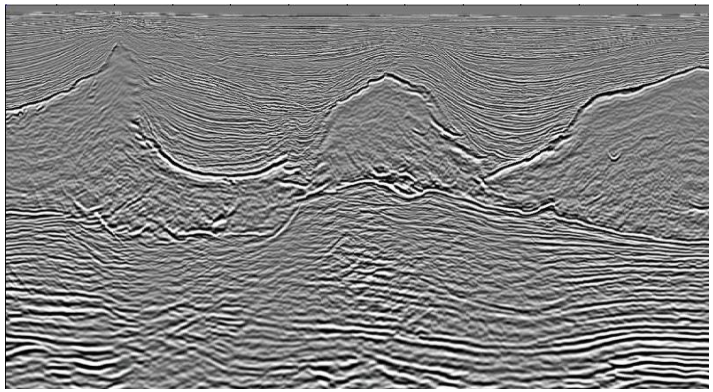








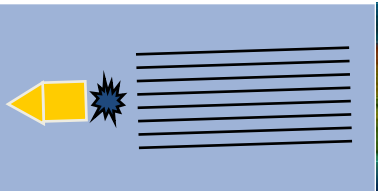
Isotropic



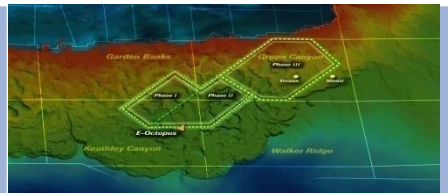
Anisotropic

MAXELER Technologies

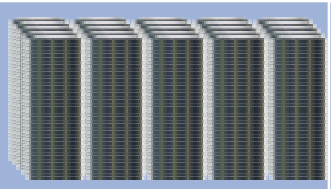
Data Rates and Computational needs



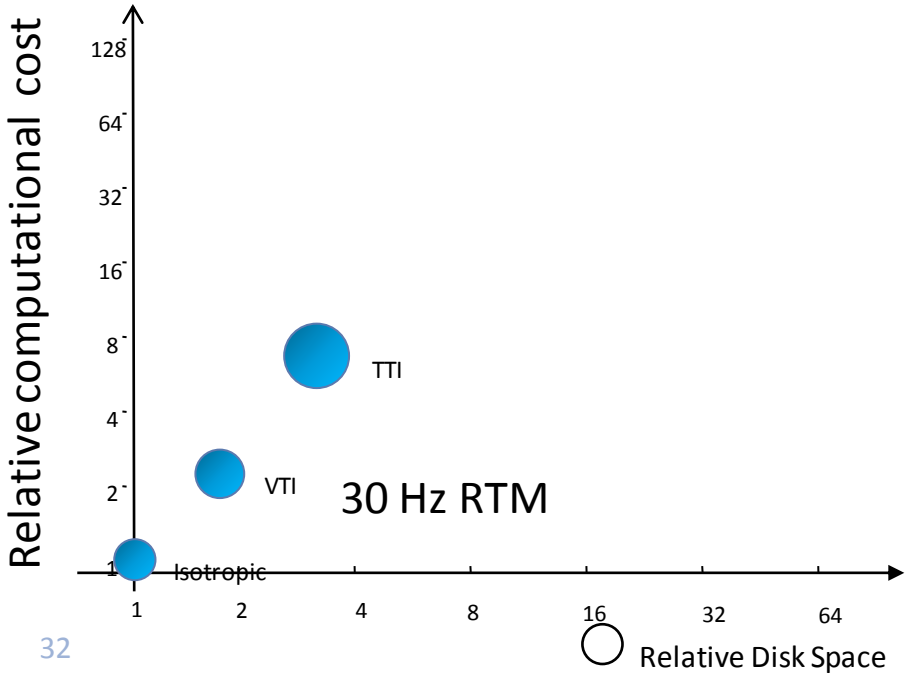
20 – 25,000 sensors
500 MB – 2 GB

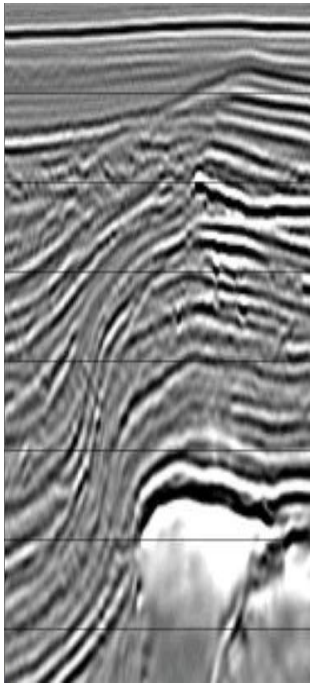


50 – 200,000 shots
50 – 200 TB Data



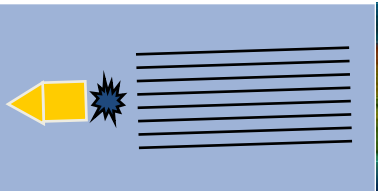
1000s node
5 – 7 days



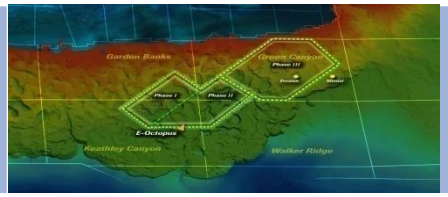


MAXELER Technologies

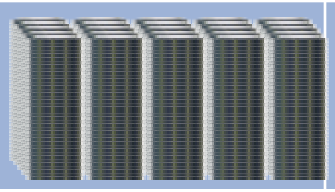
Data Rates and Computational needs



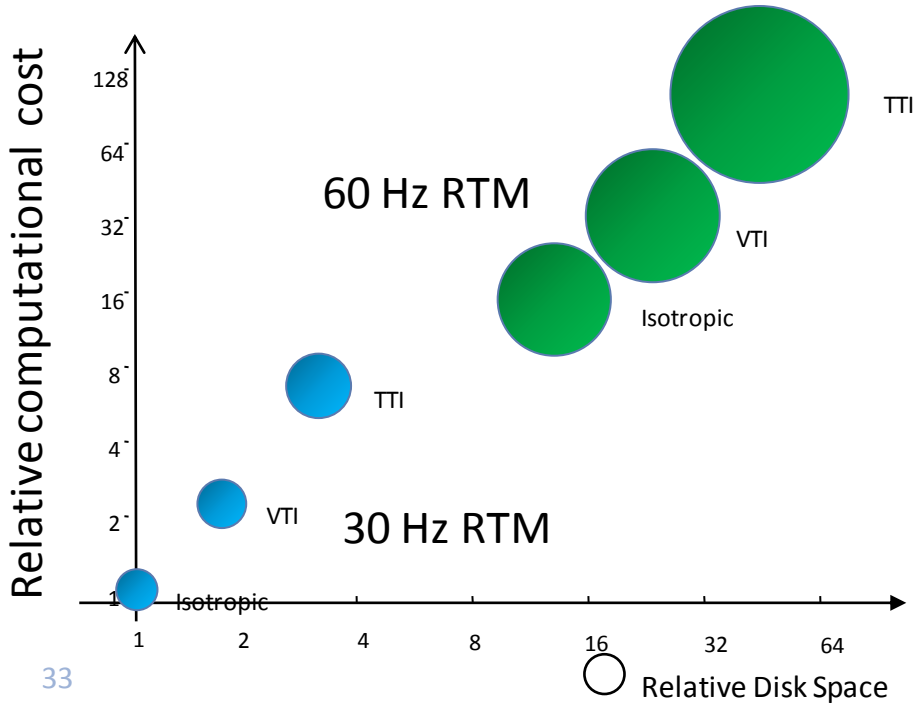
20 – 25,000 sensors
500 MB – 2 GB

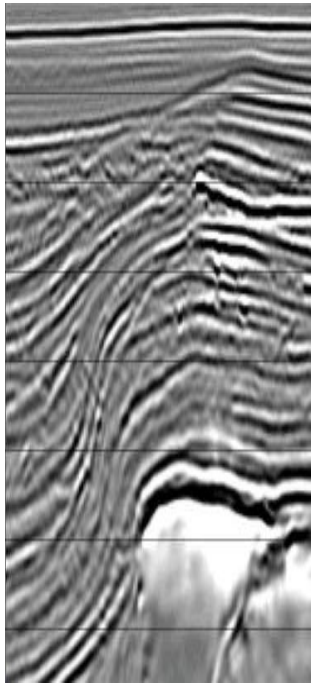


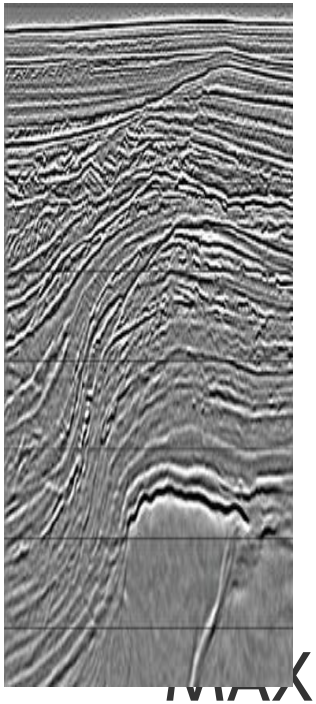
50 – 200,000 shots
50 – 200 TB Data



1000s node
5 – 7 days

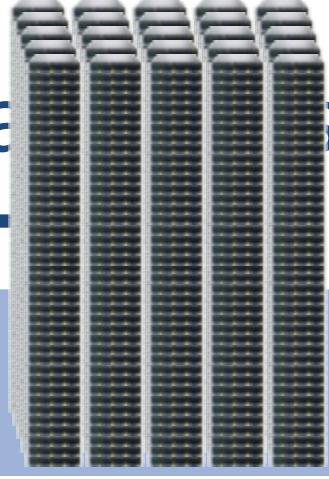


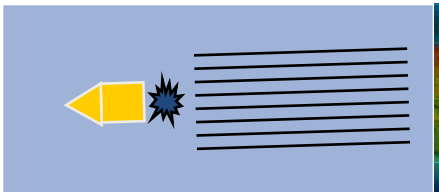


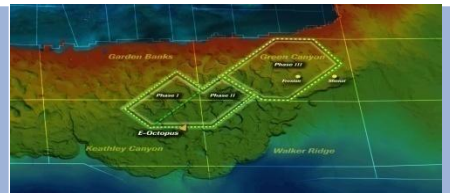


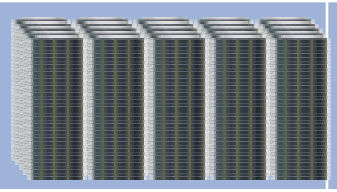
MAKELER Technologies

Data Rates and Computationa

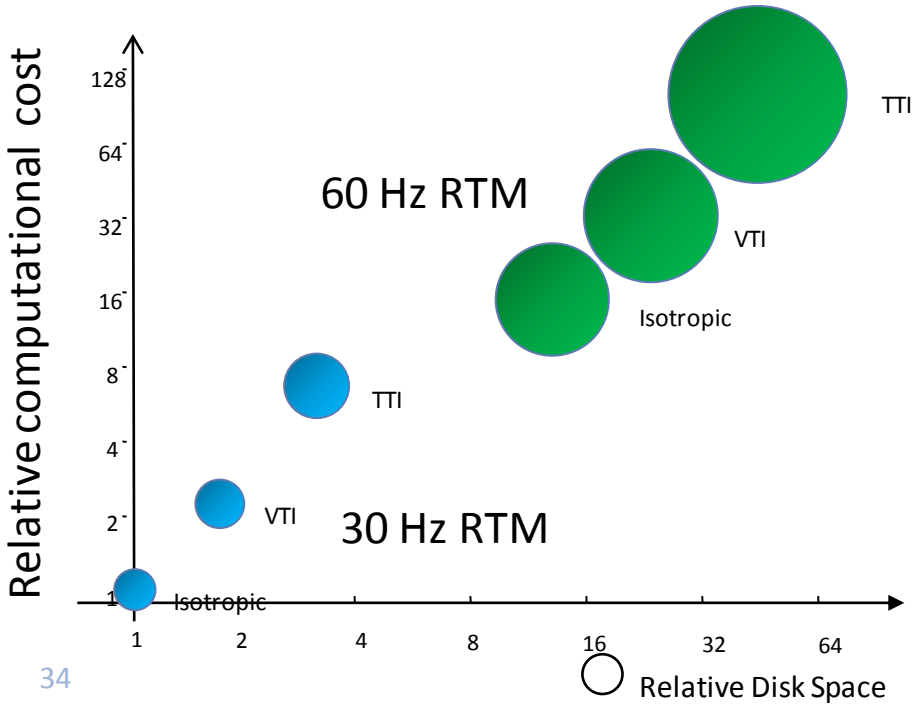


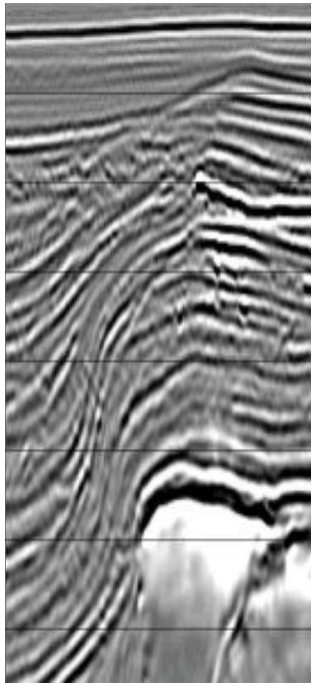


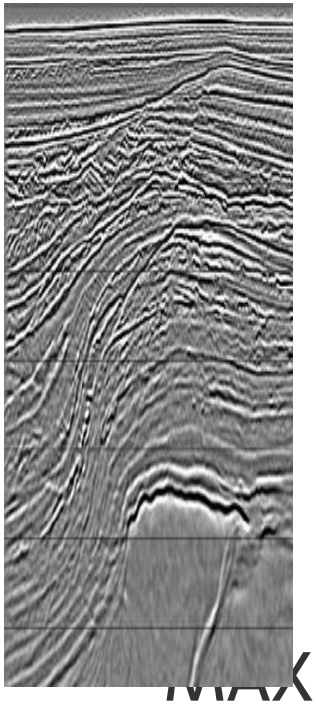




20 – 25,000 sensors 500 MB – 2 GB	50 – 200,000 shots 50 – 200 TB Data	1000s node 5 – 7 days	15 -20,000 nodes Days – weeks



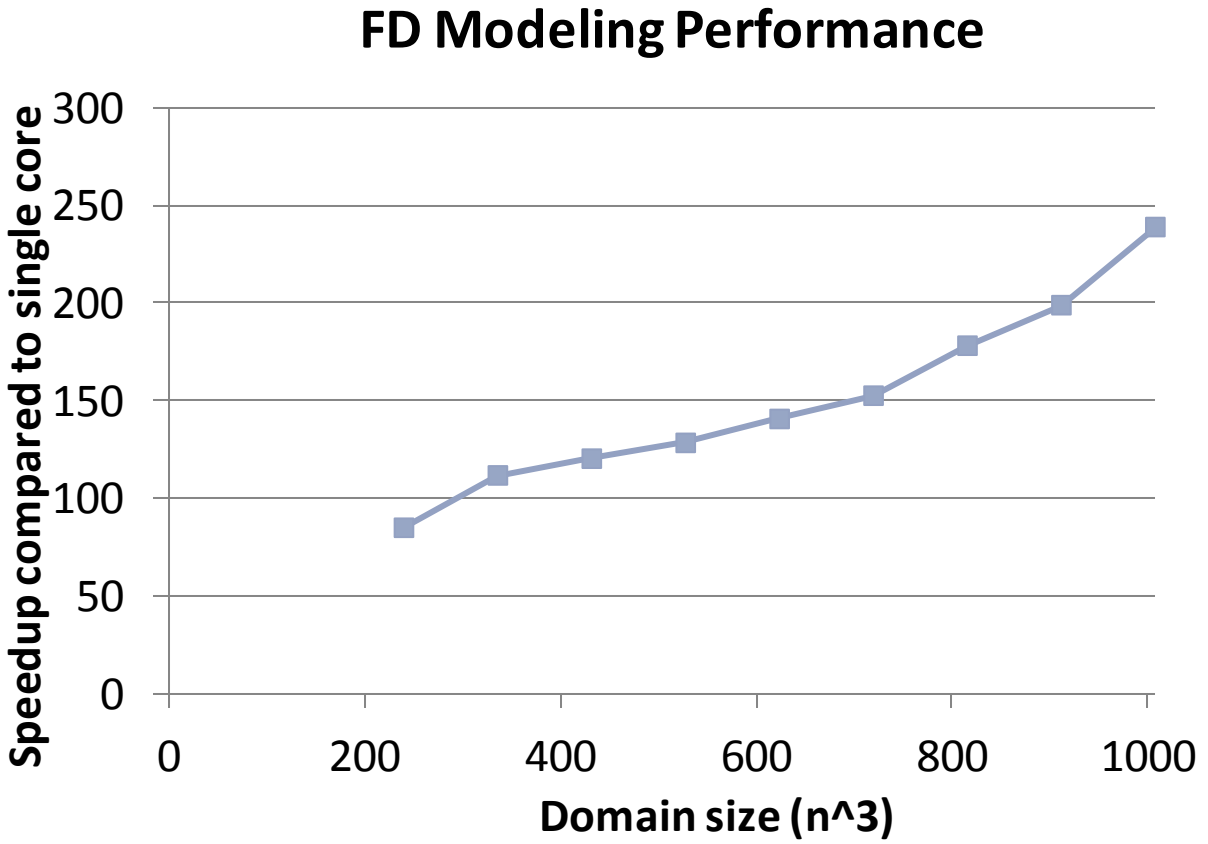




MAKELER Technologies

Accelerated Modeling

- Up to 240x speedup for 1 MAX2 card compared to single CPU core
- Speedup increases with cube size
- 1 billion point modeling domain using single FPGA card



Chevron

MAXELER Technologies

CRS Trace Stacking

P. Marchetti et al, 2010

- A ‘stack’ is a processed seismic record that contains traces that have been added together from different records to reduce noise and improve overall data quality
- CRS stacking is a data driven method to obtain a stack, based on 8 parameters
- Typical CPU runtime: 1000 cores for 1 month

eni

MAXELER Technologies

3D CRS

- Search in 8 dimensional parameter space, and evaluate result by calculating *semblance*

$$S(\vec{x}_0, t_0) = \frac{1}{M} \frac{\sum_{k=-N/2}^{N/2} \left| \sum_{i=1}^{M} a_{i,t_i+k} \right|^2}{\sum_{k=-N/2}^{N/2} \sum_{i=1}^{M} \left| a_{i,t_i+k} \right|^2}$$

- t_i comes from the CRS travel-time formula:

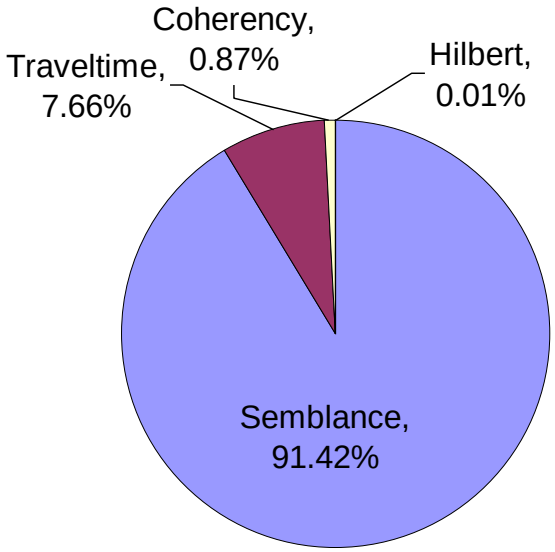
$$t_{hyp}^2 = \left(t_0 + \frac{2}{v_0} \mathbf{w}^{\mathrm{T}} \mathbf{m} \right)^2 + \frac{2t_0}{v_0} \left(\mathbf{m}^{\mathrm{T}} \mathbf{H}_{\mathbf{zy}} \mathbf{K}_{\mathbf{N}} \mathbf{H}_{\mathbf{zy}}^{\mathrm{T}} \mathbf{m} + \mathbf{h}^{\mathrm{T}} \mathbf{H}_{\mathbf{zy}} \mathbf{K}_{\mathbf{NIP}} \mathbf{H}_{\mathbf{zy}}^{\mathrm{T}} \mathbf{h} \right)$$

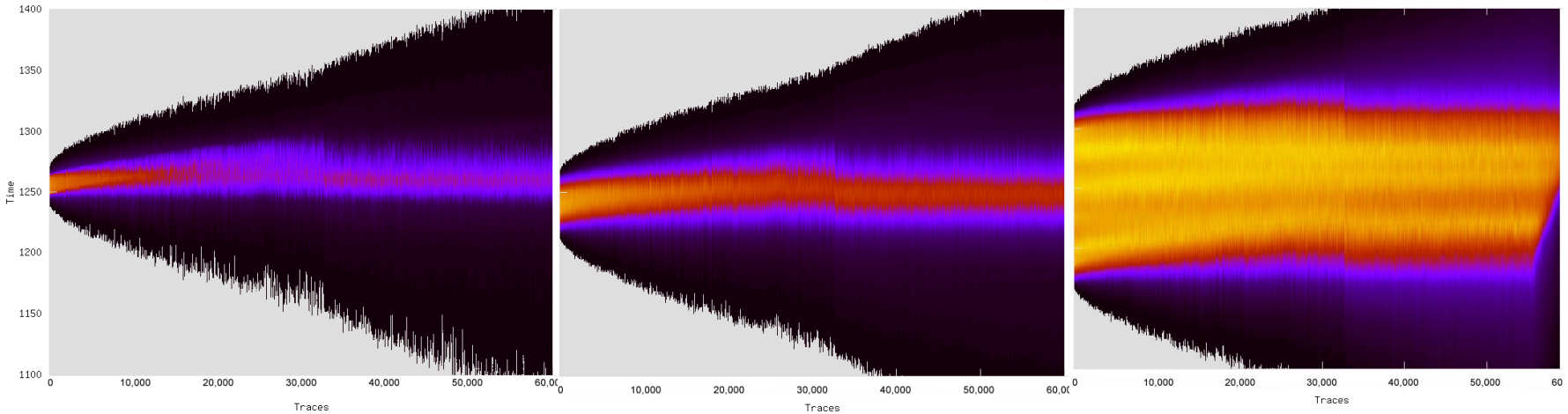
eni

MAXELER Technologies

CRS Application Analysis

- Runtime dominated by travel time and semblance calculation
- CPU: compute samples in series
- FPGA: compute multiple samples in parallel





1 t_0 sample

16 t_0 samples

64 t_0 samples

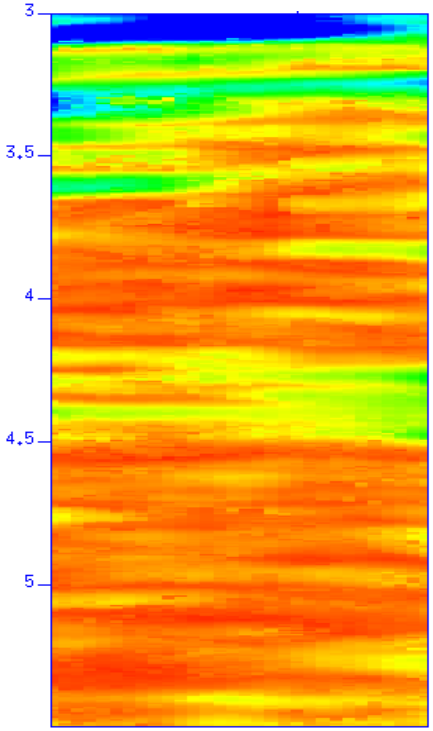
eni

MAXELER Technologies

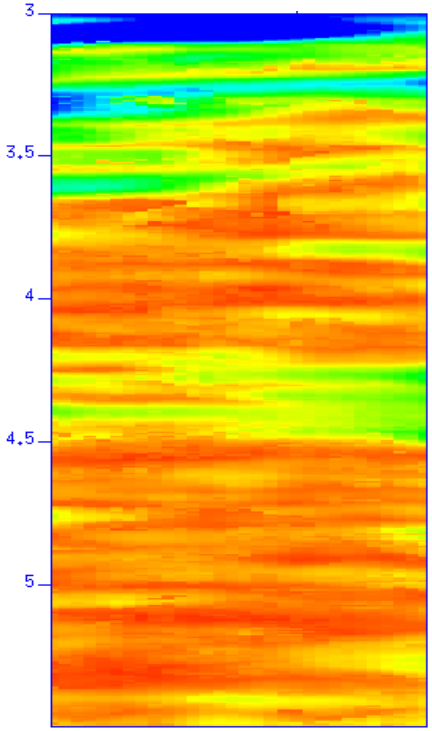
Accelerated CRS

- Performance of one MAX2 card vs. 1 CPU core
  - Land case (8 params), speedup of 230x
  - Marine case (6 params), speedup of 190x

CPU Coherency



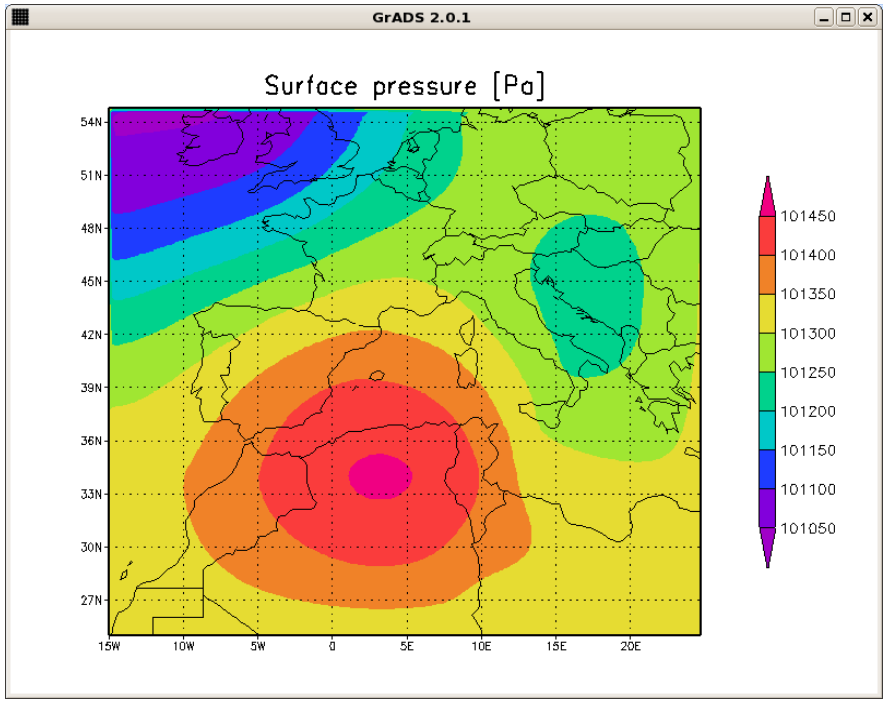
MAX2 Coherency

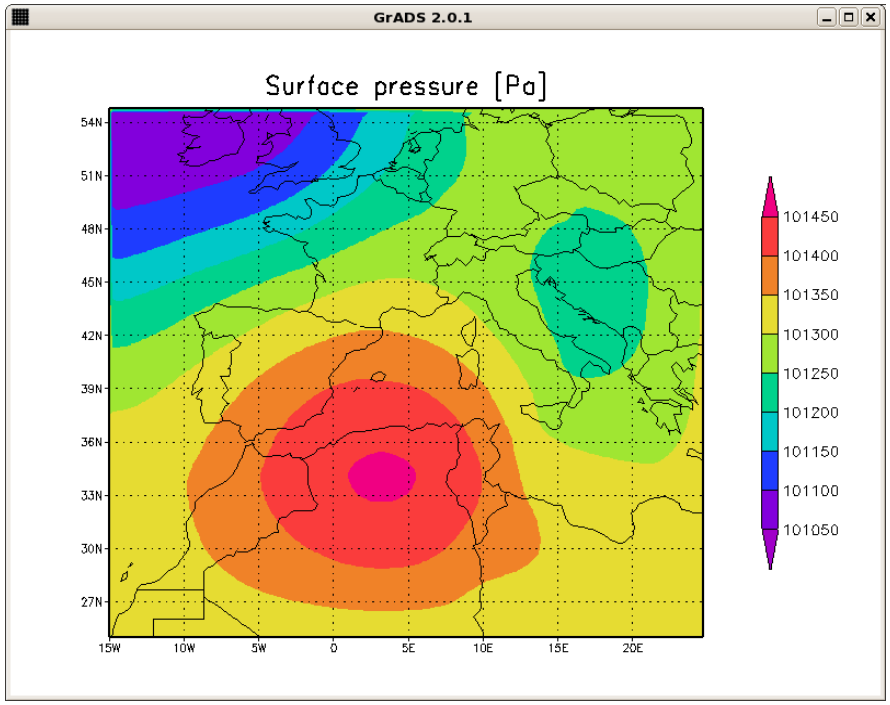


eni

MAXELER Technologies

Meteo Project in Sardinia





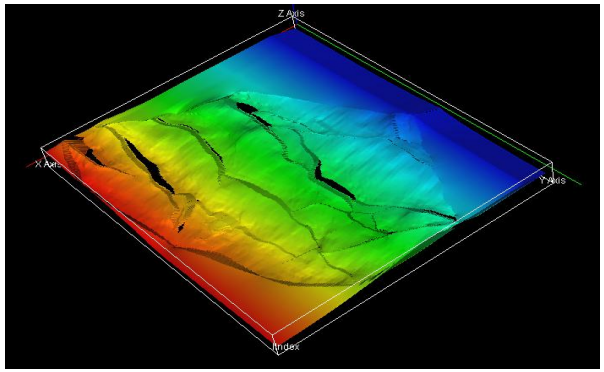
1U CPU Node
Wall Clock Time: 2 hours

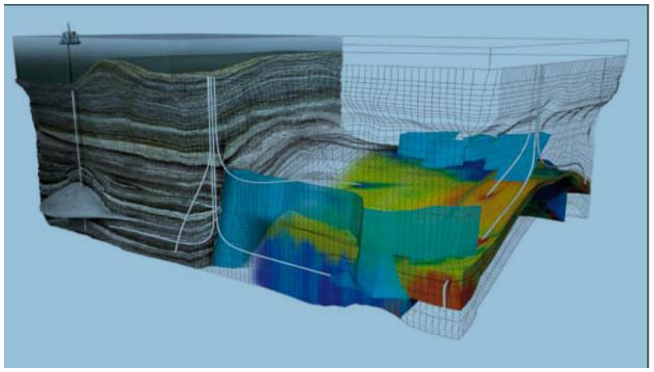
1U Dataflow Node
less than 2 minutes

Problem size: (Longitude) 13,600 Km x (Latitude) 3330 Km
Simulation of baroclinic instability after 500 time steps.

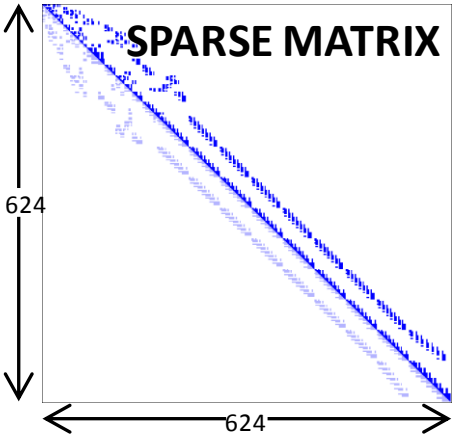
Geomechanics Simulation

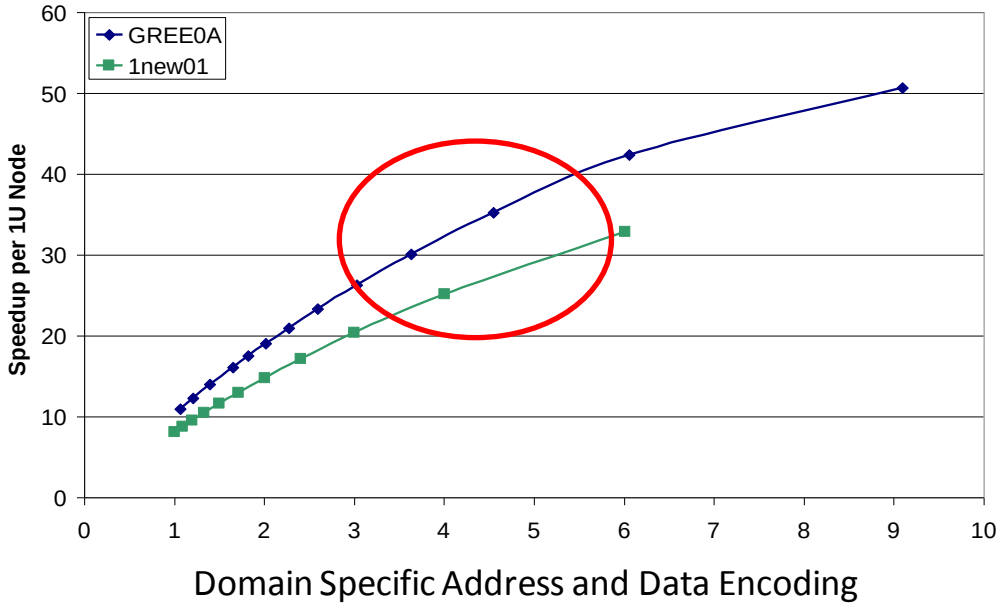
Schlumberger





1 MaxNode achieved 20-40x vs CPU node.



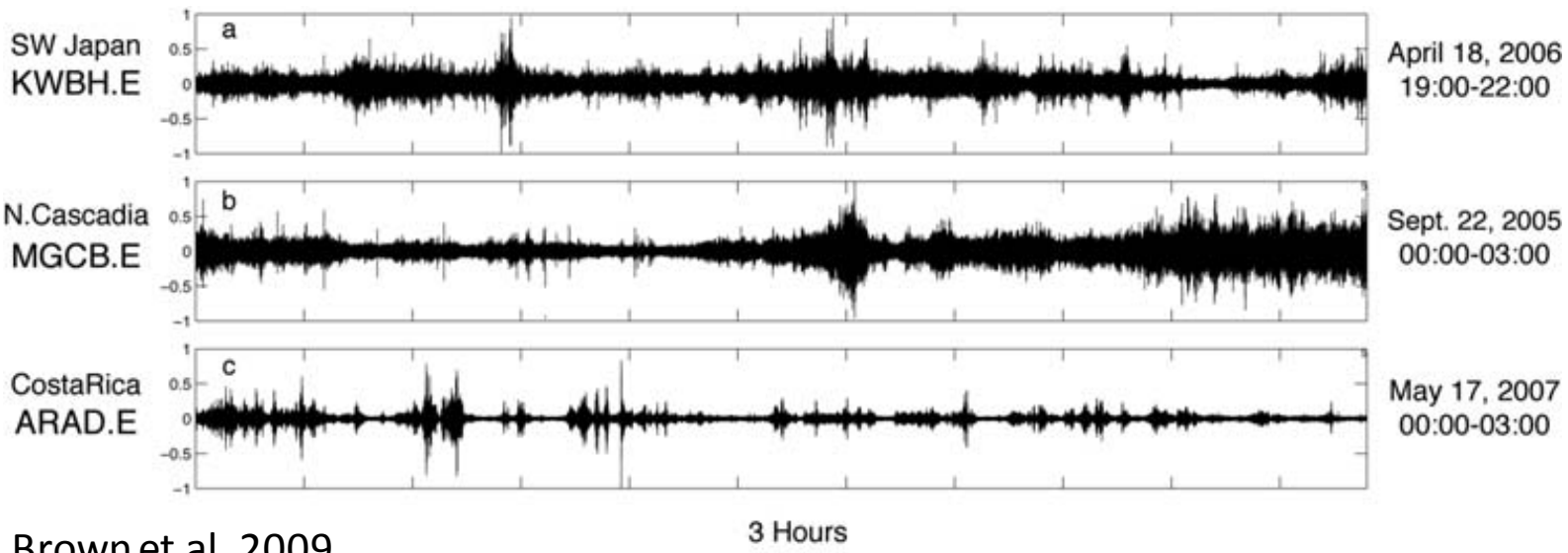


HotChips Conference, Stanford, 2010

MAXELER Technologies

Low frequency Tremor Detection

(joint work with Robert G Clapp at Stanford)



Brown et al. 2009

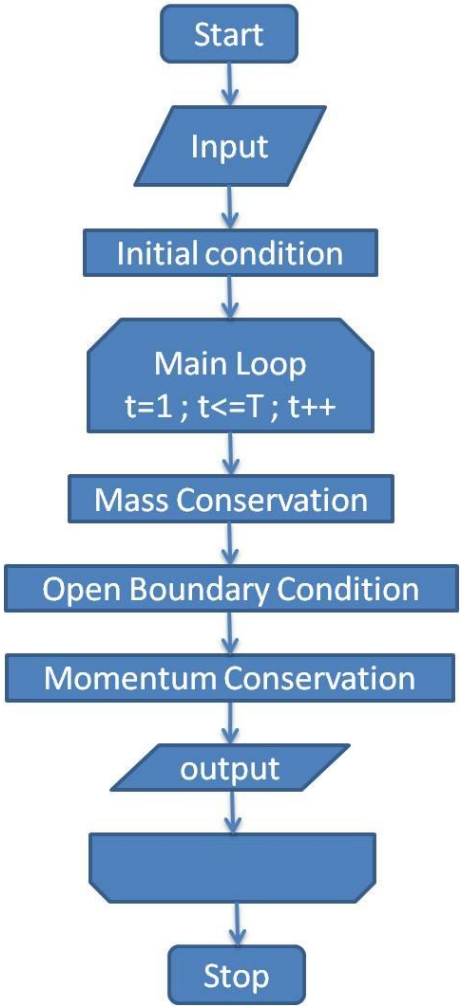
$$M(ia, ib) = \sum_{is=0}^{ns} \sum_{it=0}^{nw} u(it + ia, i2)u(it + ib, is)$$

Where u(it,is) is a collection of seismographs, nw - is the correlation window, M(ia,ib) is the array describing shifts. High values of M => more tremor events

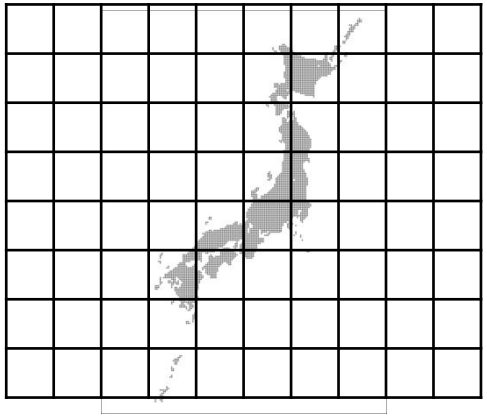
MAXELER Technologies

TUNAMI simulation

joint work with University of Tokyo



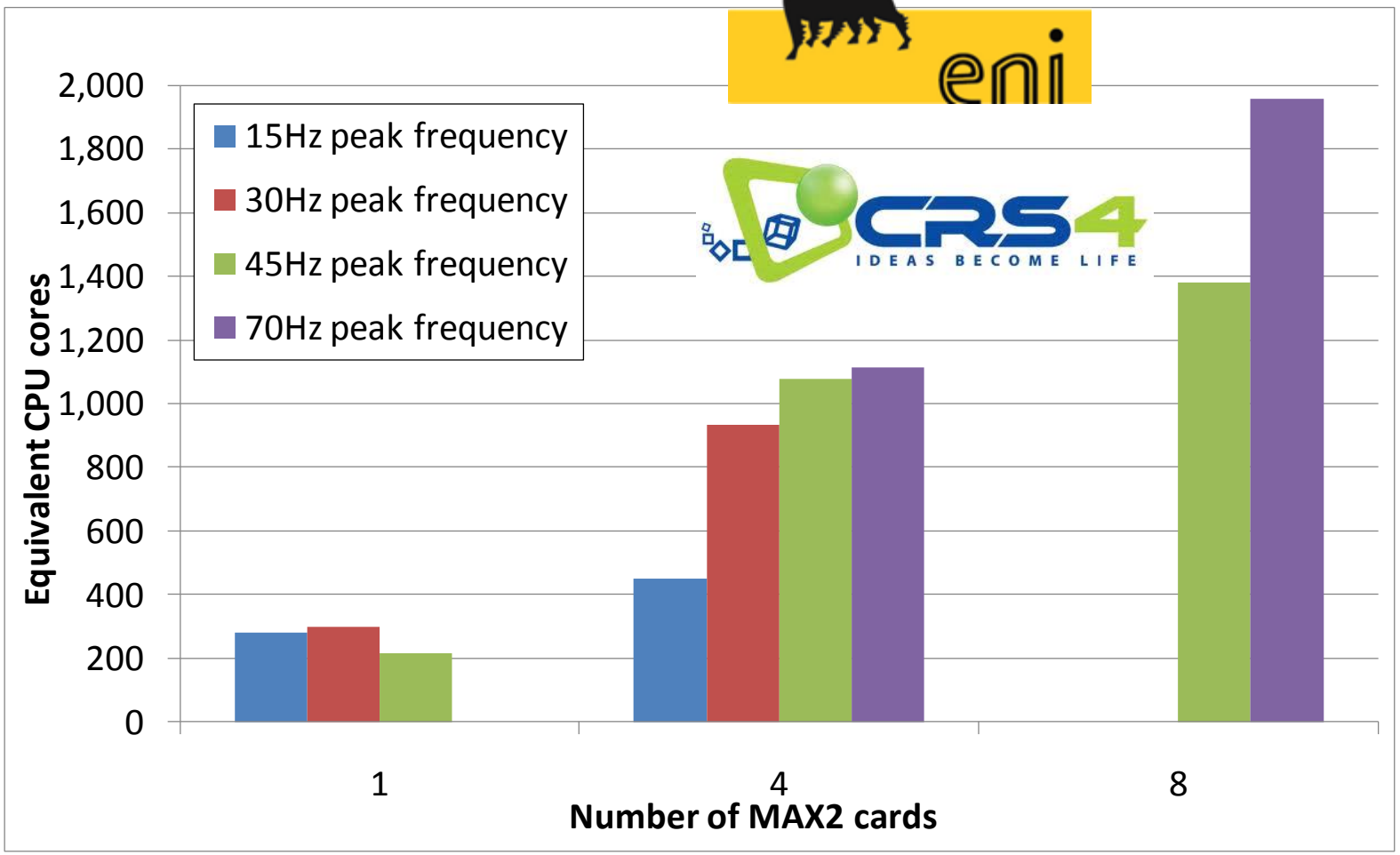
- Input
  - Still Water Depth
  - Origin of Epicenter
- Output
  - Wave Surface Level of each time step



MAXELER Technologies

3D Finite Difference as reported by ENI

2 MaxNodes = 1,900 Intel CPU cores, published at SEG 2010



Compared to 32 3GHz x86 cores parallelized using MPI

MAXELER Technologies

Computational Finance

- Compute value of complex financial products
- Compute risk: How sensitive is price to moves in the market?
- Typically computed overnight on hundreds to thousands of CPU cores. But:
  - Looking at yesterday's risk is like driving by looking in the rear-view mirror
  - We really need to evaluate scenarios: *what happens if Greece leaves Euro?*
  - Regulatory pressure for more and better analytics

Maxeler-JP Morgan Collaboration

- Winner of Waters Technology AFTA award for *Most Cutting-Edge IT Initiative* in December 2011
- Three-year project with bank's applied analytics group, working on production code.
- Code speedups versus single-core of Xeon CPU:
  - Multi-factor Monte Carlo interest rate models: 284x
  - Multi-nominal tree models: 207x
  - Base correlation with stochastic recovery: 155x
  - (Information published in AFTA award citation)
- *Target for the first quarter of this year is to achieve sub-one-hour for calculating the risk across JP Morgan's entire, highly complex, long-dated forex book which at the moment takes seven to eight hours*
  - Stephen Weston in Risk Magazine: 6 Feb 2012

MAXELER Technologies

Summary & Conclusions

- Forget benchmarks: Focus on solving real-world problems
- Dataflow supercomputing can
  - Maximize performance
  - Minimize space and power consumption

Things We Didn't Talk About

- Chips designed for Dataflow Computing
- How to analyze and re-architect an application for Dataflow
- How to implement an application on Dataflow computers
- Computer Arithmetic options and tradeoffs
- Low-latency computing in Financial Exchanges
- Genomics and Biological Modelling