

Building massive-scale super-computing systems in the Cloud



Tobias Becker

21 Jan 2019, HiPEAC Workshop on Reconfigurable Computing, Valencia

A perspective of the recent past



So let's go back in time...

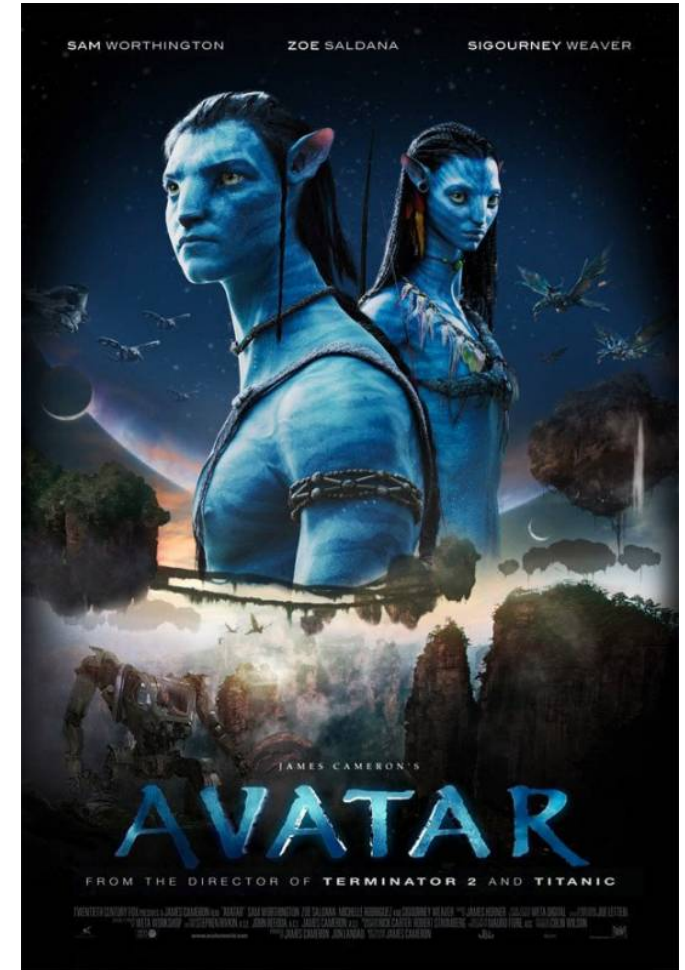
The year 1999...



Intel Pentium III
1 core, 450MHz,
250 nm process



The year 2009...



**Core i7 (1st gen),
4 cores, 3GHz,
45nm process**



The year 2019 – compared to the past

Pentium III

1999

1 core

450 MHz

250 nm



Core i7 (1st gen)

2009

4 cores

3 GHz

45 nm



Core i7 (9th gen)

2019

6-8 cores

3.2-3.6 GHz

14 nm



The rise of the accelerators



2008: Acoustic wave modeling (MAX2)



2011: Credit derivatives modeling (MAX3)



2014: Real time risk (MAX4)  **CME Group**

2018: QCD, QuantumEspresso, SpecFEM3D, Nemo (MAX5)



2015: **Intel Agrees to Buy Altera for \$16.7 Billion**

On-again-off-again deal is latest acquisition in active semiconductor sector

2016: **MICROSOFT BETS ITS FUTURE ON A REPROGRAMMABLE COMPUTER CHIP**

Amazon's Xilinx FPGA Cloud: Why This May Be A Significant Milestone

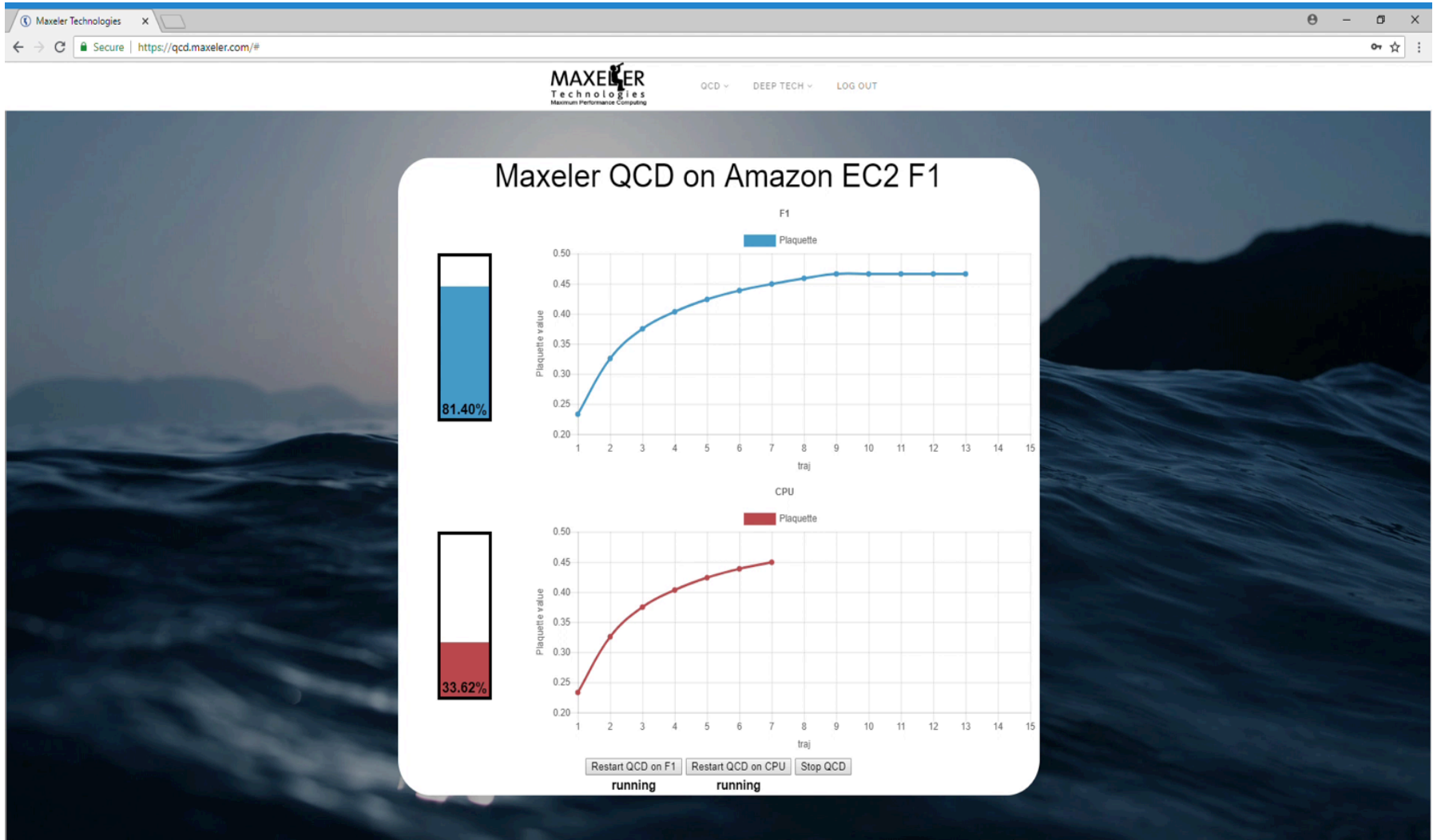
powered by



2018: Xilinx launches Alveo datacenter card



Maxeler QCD on Amazon EC2 F1



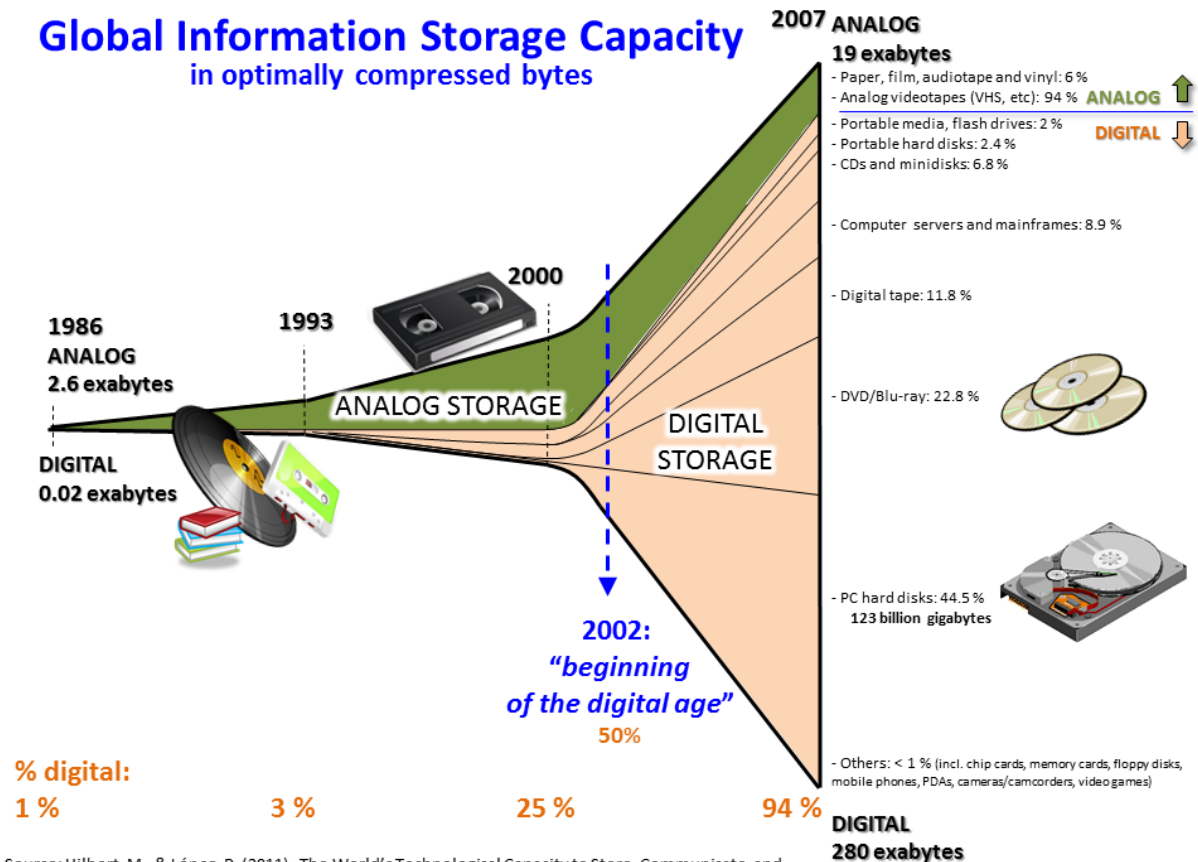
The rise of data

2007:

mobile and pervasive computing driving massive expansion of data



Global Information Storage Capacity in optimally compressed bytes

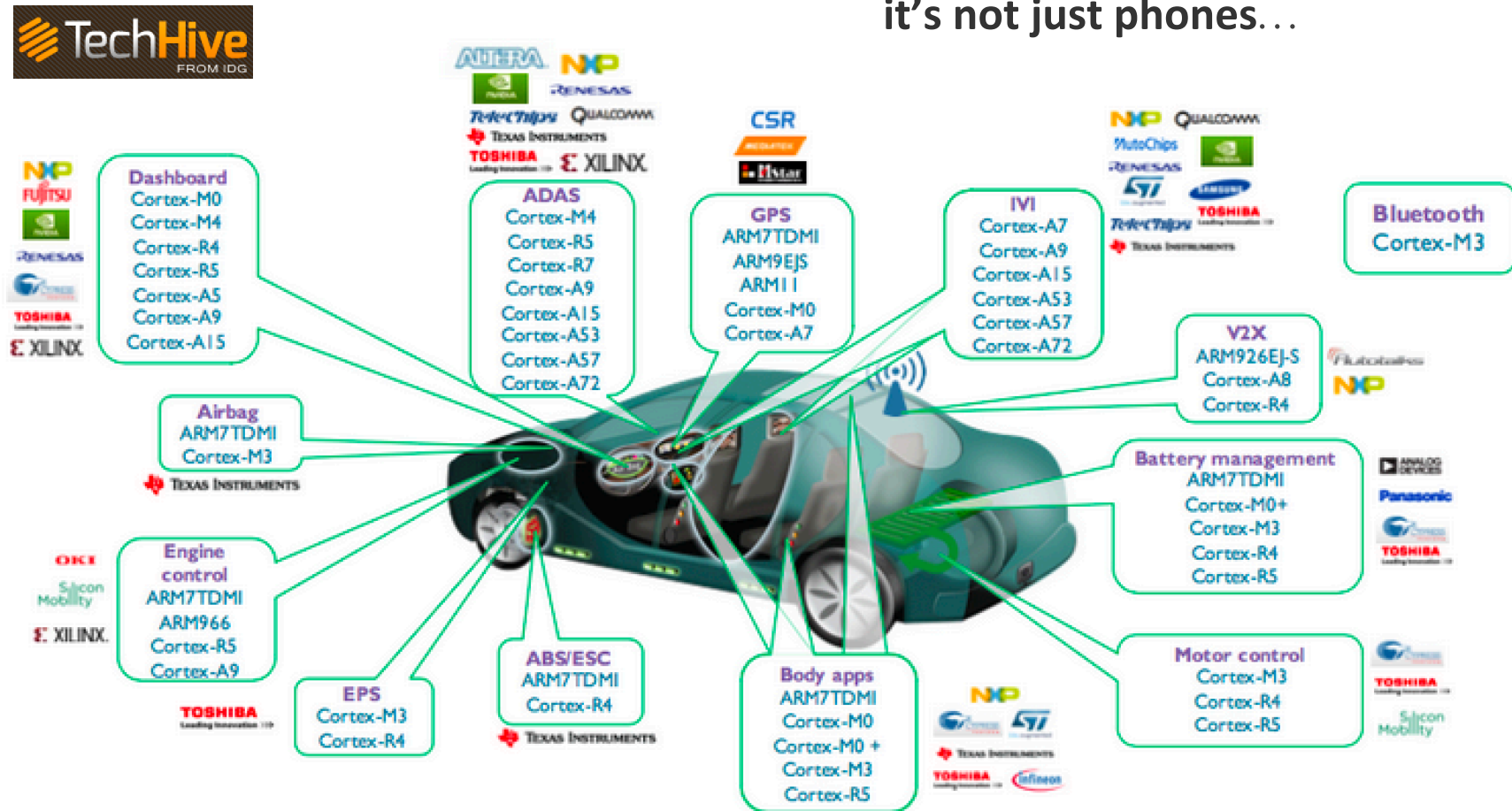


Source: Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, 332(6025), 60–65. <http://www.martinhilbert.net/WorldInfoCapacity.html>

ARMageddon?

it's not just phones...

ARM



©ARM 2015

and just imagine the software running on all these processors

No surprise modern cars are unreliable as never before

The future challenge: MW, not FLOPS

'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

US edition ▾
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."

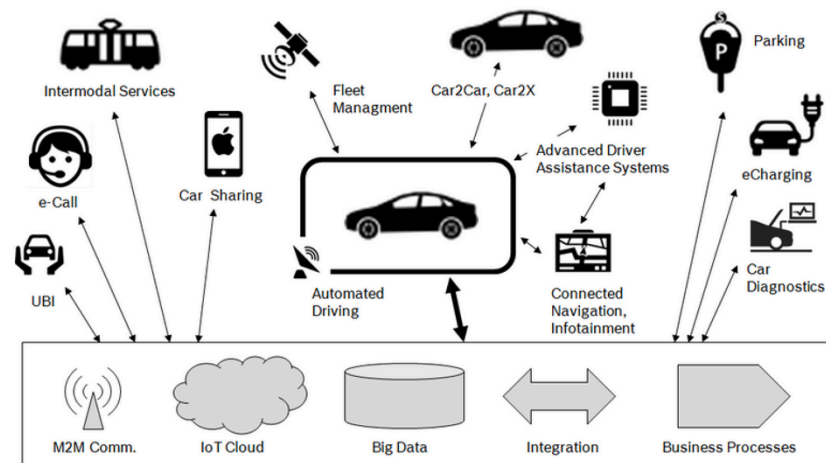
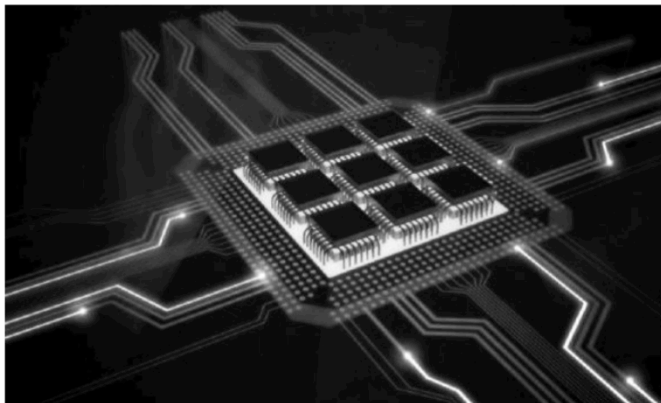
How dataflow computing can help

- Efficient data processing at massive scale requires radical change of thinking
 - **Data movements are the main contributor to energy consumption**
- System-wide design works better than focusing at the back end only
 - **Doing just the required task at the right place reduces energy wastage**
 - It also helps with redundant data and keeps only meaningful information
- **Build direct path from problem to solution** (both top-down as well as bottom-up)
 - Balance between reductionism and holism while benefiting from specialisation
- Focus on the data and its movements at all system levels
 - **Computation happens as a side effect** as the data moves through the system
- **Analyze → Architect & Model → Build → Deploy** is the key

INTEL'S EXASCALE DATAFLOW ENGINE DROPS X86 AND VON NEUMANN

August 30, 2018 Timothy Prickett Morgan

 THE NEXT PLATFORM

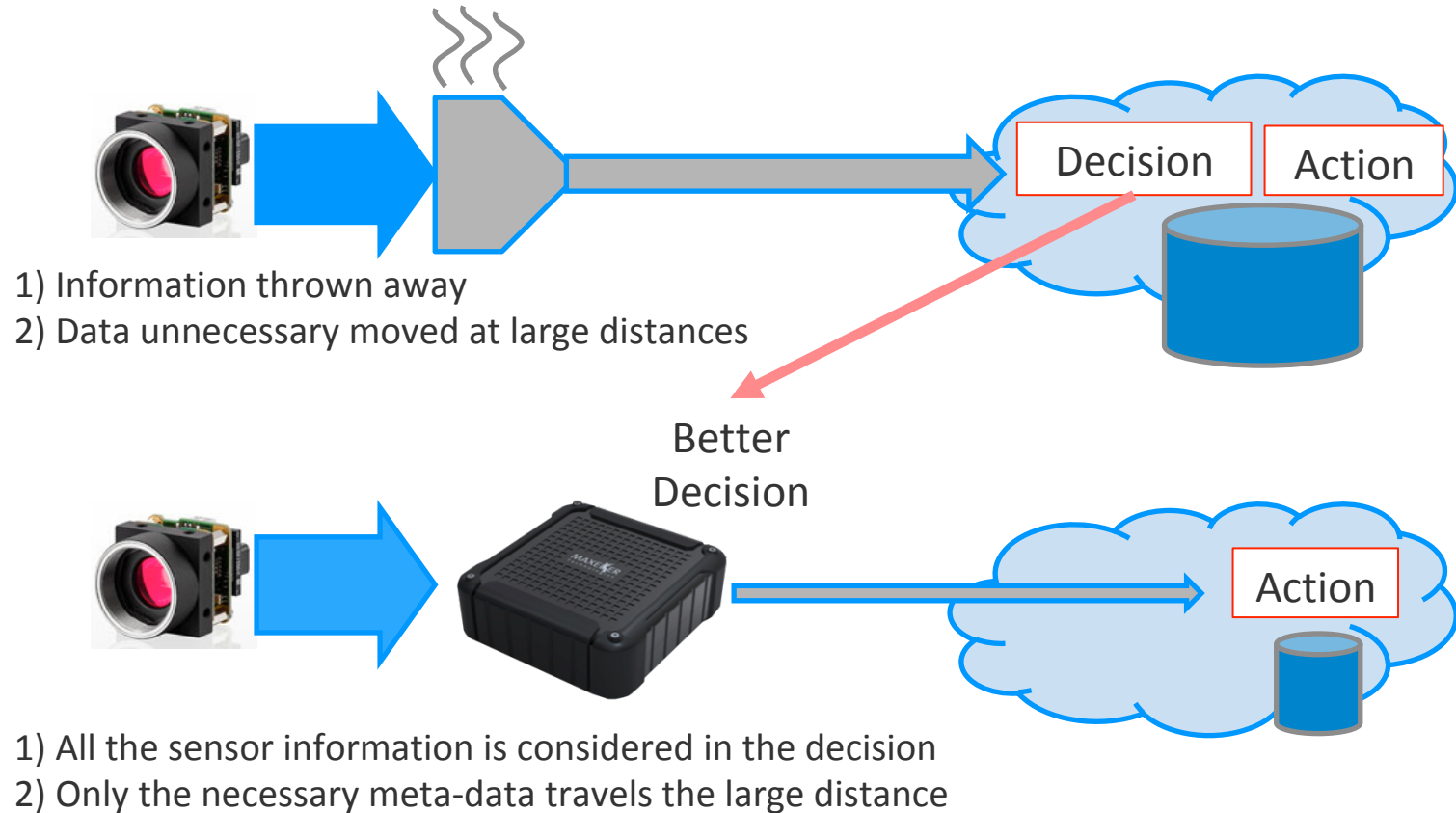


Source: www.enterprise-iot.org

ENTERPRISE  IoT

Motivational Example: Hyperspectral Camera

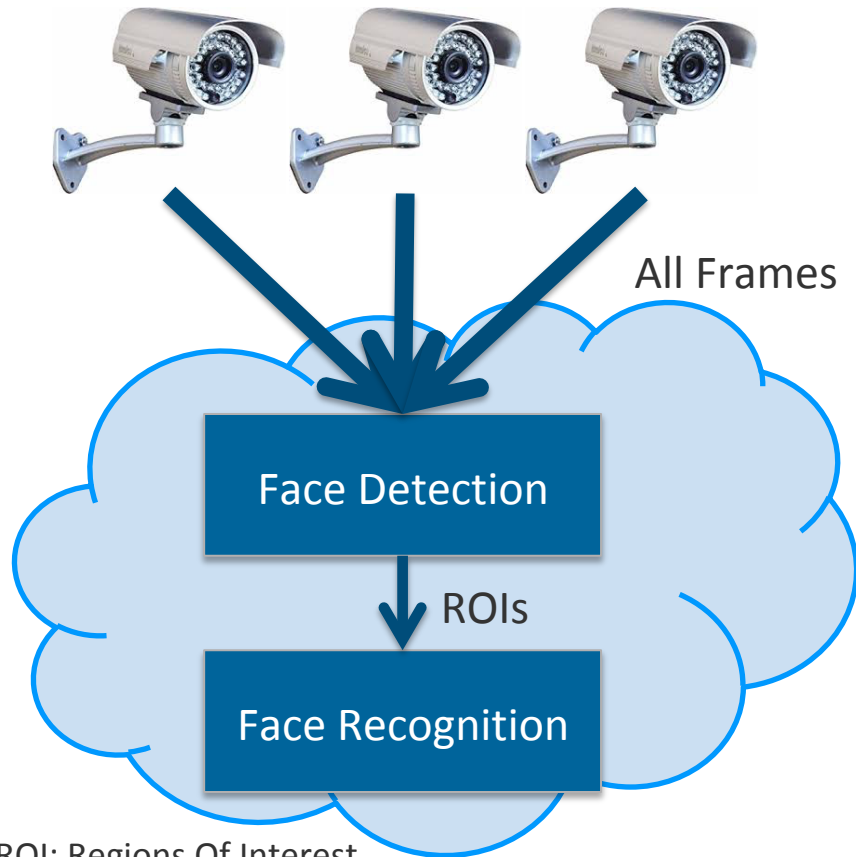
Bringing decisions closer to the data source for low latency responses



Boost the overall application quality of service while minimizing data movements

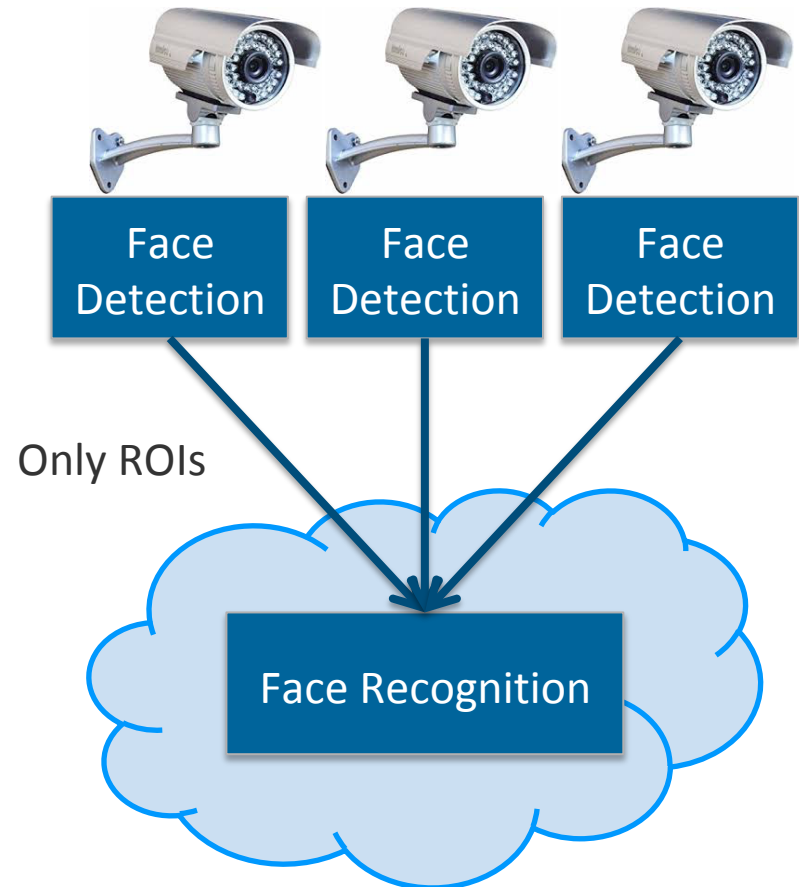
Motivational Example 2: Surveillance

Monolithic



* ROI: Regions Of Interest

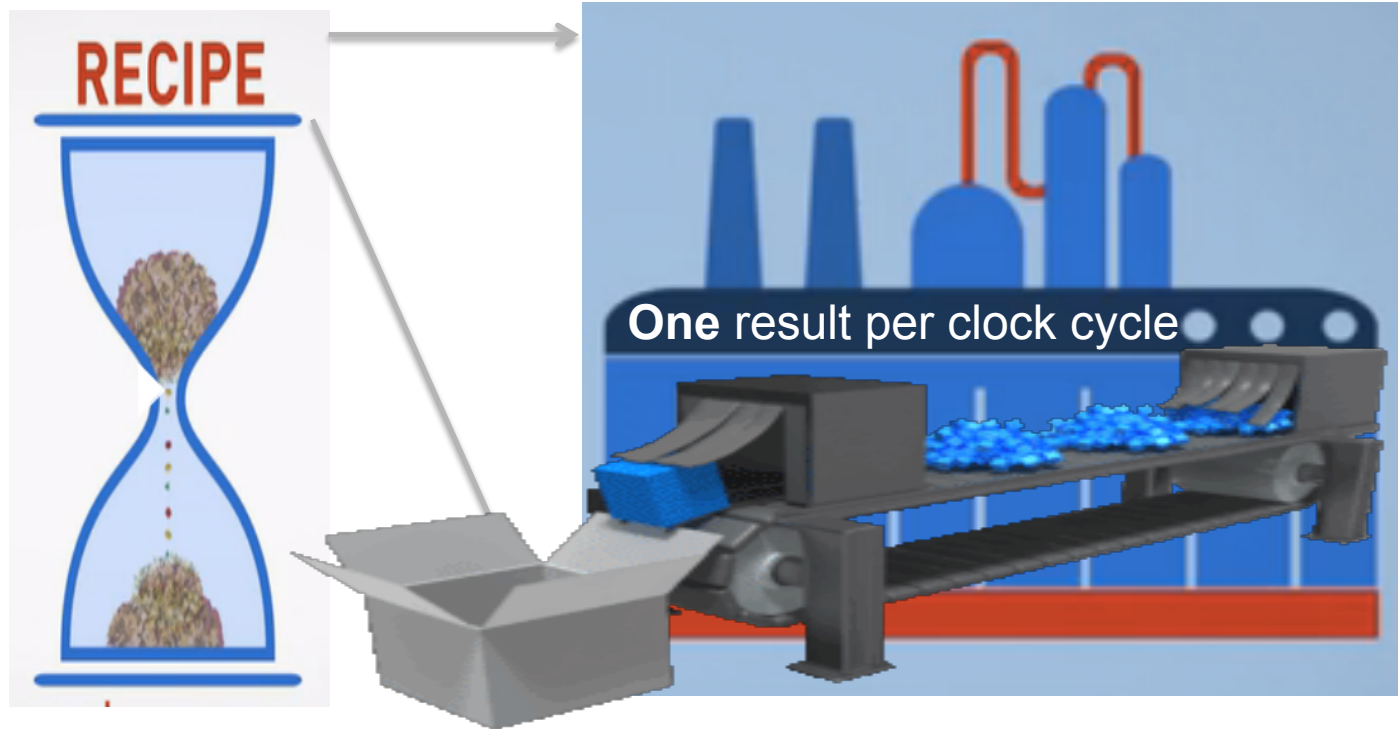
Well balanced, Cumulonimbus system



Interestingly these are pretty invariant algorithms that can be “hard wired”

Efficiency through Computing in Space

Treat computation as data “mass production” problem

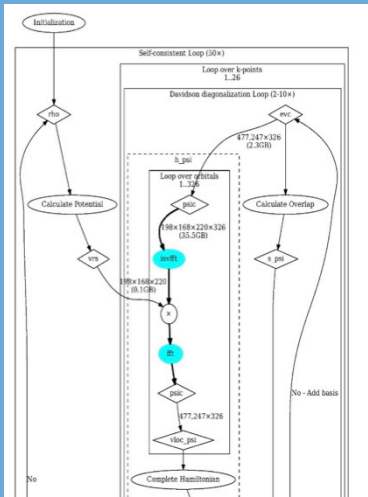


Computing in Time:
*Follow a recipe step by
step one at the time*

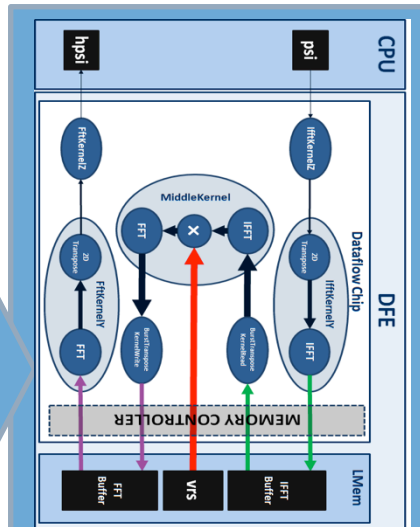
Computing in Space:
*Build a “recipe specific” factory with
multiple paths performed simultaneously*

Efficient, predictable, reliable mass production of data

Computing in Space: Design process



LOOP FLOW GRAPH



ARCHITECTURE

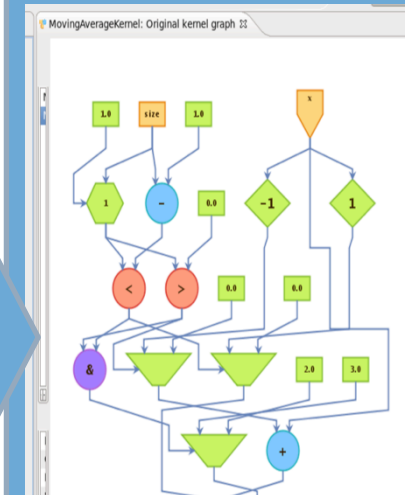
```
class MovingAveKernel extends Kernel
{
    MovingAveKernel(KernelParameters
        super(parameters);

    DFEVar x = io.input("x", dfe);

    DFEVar prev = stream.offset(
    DFEVar next = stream.offset(
    DFEVar sum = prev + x + next;
    DFEVar result = sum / 3;

    io.output("y", result, dfe);
}
```

MaxJ

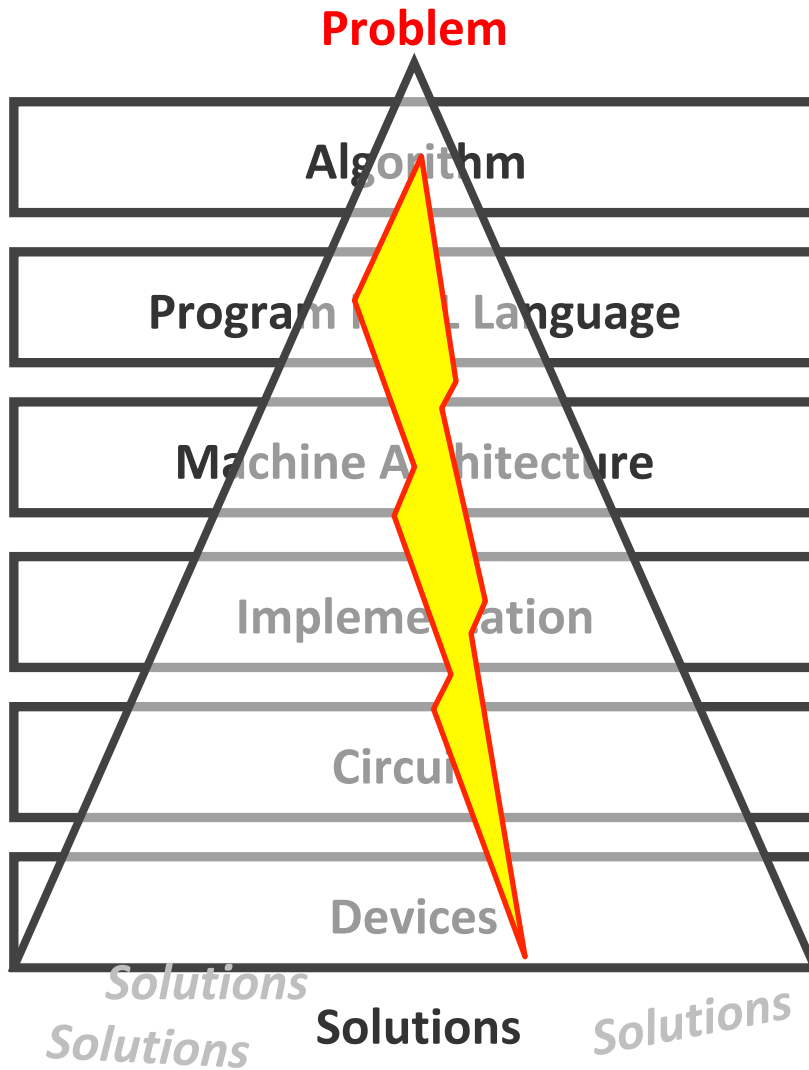


MULTISCALE DATAFLOW

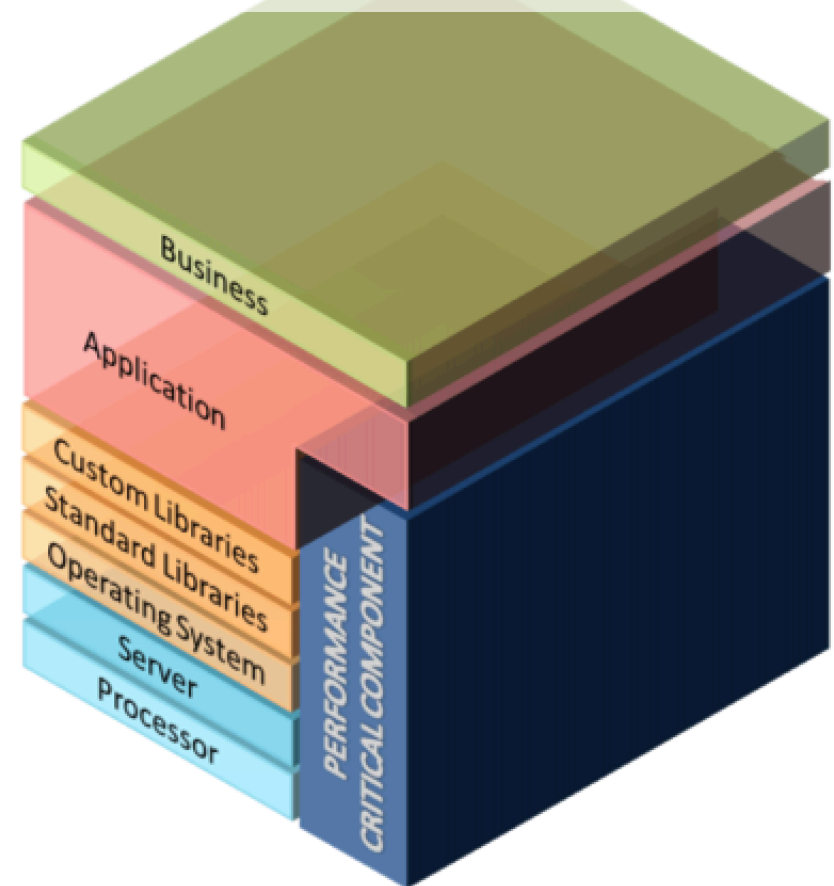
MaxJ SIMULATOR
Debug and Compare to C-models

Proven datacentre technology, ready for the Cloud

Thinking Vertically about Computing Problems



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



Maxeler's Dataflow Engines (DFEs) based technology facilitates the above

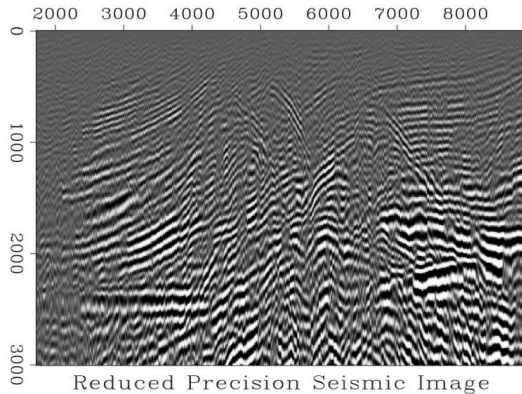
Optimizations at all levels

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	=> create the illusion of '0' and '1'

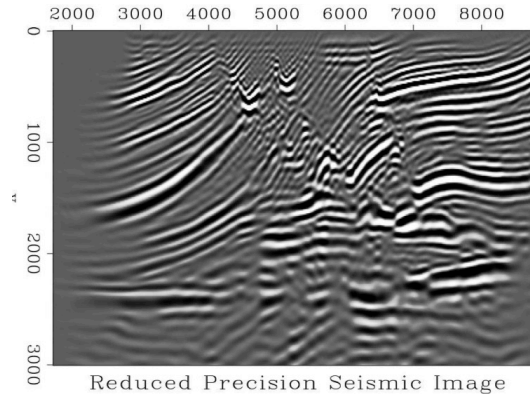
and more, e.g., trade Communication (Time) for Computation (Space)

Example: Exploring fixed-point numerics

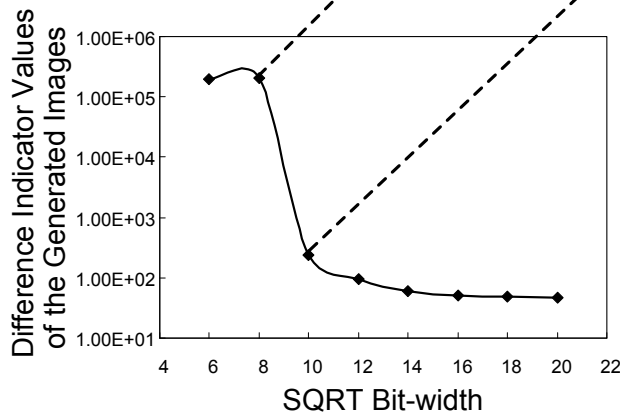
8-bit fixed-point



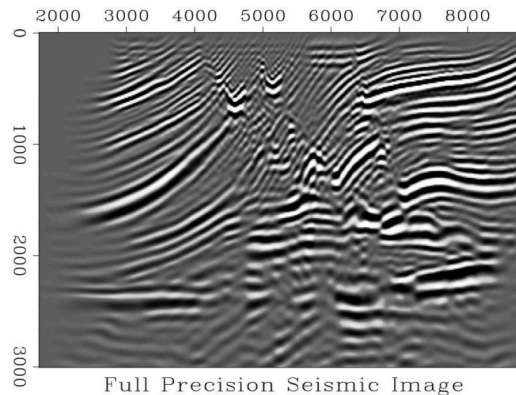
10-bit fixed-point



Different parts are explored *separately*, i.e., when we investigate one part, we keep the bit-widths in other parts a constant high value



'true' image: single-precision floating-point



Similarly, we observe a significant drop of the error when the SQRT bit-width increases from 8 to 10

Similar precision thresholds observed in both synthetic and field results. This behavior enables an automatic tool to determine the minimum precision that still keeps the result *good enough*

[H.Fu, W. Osborne, R. G. Clapp, O. Pell, Accelerating Seismic Computations [], 70th EAGE Conference, Italy, 2008]

Slotnick's law: We have to do a bit of work

Slotnick's law (of effort):

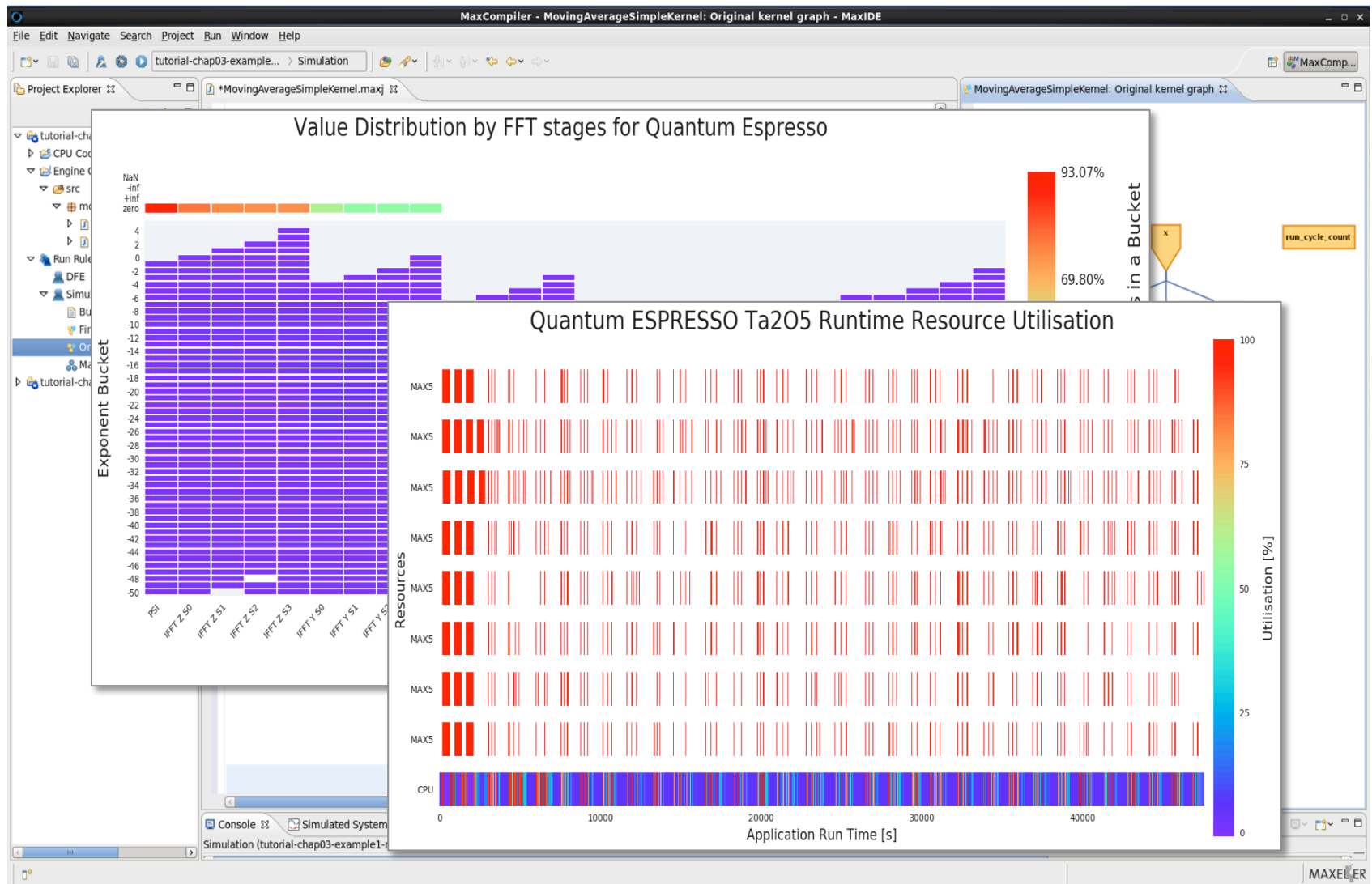
“The parallel approach to computing does require that some **original *thinking*** be done **about numerical analysis and data management** in order to secure efficient use.

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



With the right tools, we can do it!



Some positive observations

- This approach helps cope with “feature creep”
 - Minor changes can have big impact
 - Every change needs careful motivation and is a trade-off
- The system-wide dataflow model brings new insides
 - Easier to translate basic units / costs to business value
- It is definitely not easier than just writing code
 - Requires **thinking**, deep understanding and discipline
 - But this can be a strength
- It is the only way to “ensure efficient use”
 - And can bring a lot of satisfaction and fun
 - You may need to “unlearn things you have learned”

Challenges for dataflow computing

1. Higher chip / system price compared to microprocessors
2. Lead design times (3 months in the best case)
 - a. Complex numerical transformations
 - b. Non-trivial area and data movement optimizations
3. “Painfully slow” Place & Route times (24+ hours)
4. Current reconfigurable HW technology is sub-optimal
 - On-chip memory not built for stream processing
 - On-chip interconnect overdesigned for Dataflow
 - Need to compete at 200MHz with processors at 3GHz
5. Designer productivity should improve (Tools and Methods)

Conclusion

- How do we build future massive scale systems?
- Let's re-think systems from the ground up!
- We need to challenge architecture, conventional solutions, languages, and software engineering processes!
- Focus on data movement first!
- We are ready to build large-scale computer systems for the next 10 years and beyond!



Some links with more information

Maxeler Multiscale Dataflow Computing:

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

Computing in Space explained by Mike Flynn:

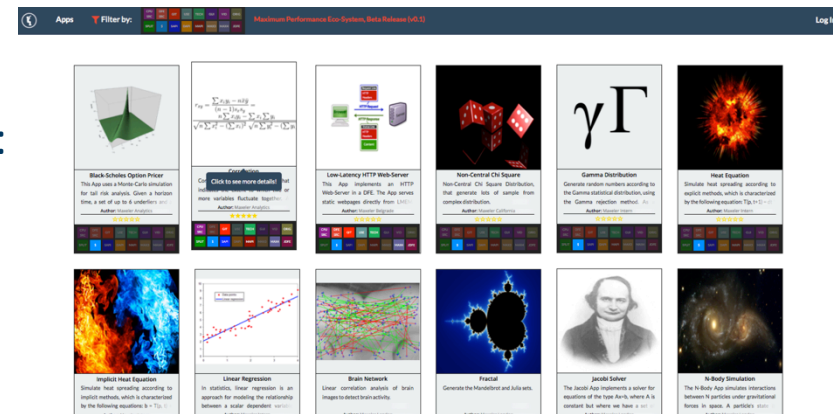
<http://www.openspl.org/what-is-openspl/>

Computing in Space Course at Imperial College:

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

Exciting Applications for DFEs (and JDFEs):

<http://appgallery.maxeler.com>



Juniper Networks' DFEs (JDFEs):

<http://investor.juniper.net/investor-relations/press-releases/press-release-details/2015/Juniper-Networks-Advances-Application-Performance-and-Acceleration-with-New-Compact-Compute-Integrated-Network-Switch/default.aspx>

NEW YORK—(BUSINESS WIRE)— Juniper Networks (NYSE:JNPR), the industry leader in network innovation, today unveiled the QFX5100-AA, a new application acceleration switch, and the QFX-PFA, a new packet flow accelerator module, together providing a ground-breaking data center switch that consolidates compute resources and customizable logic into the network. Building on Juniper's proven QFX product family, the QFX5100-AA and QFX-PFA in partnership with Maxeler Technologies' customizable software logic will significantly accelerate business-critical applications in latency-sensitive computing environments. Ideal for financial services institutions, the switch and module will vastly improve the performance of applications that handle market data feeds, social media feeds, order execution routing and big data analytics, allowing financial firms to make decisions faster and take advantage of real-time market opportunities.

The Maxeler Dataflow Software Eco-system

Apps Filter by: Maximum Performance Eco-System, Beta Release (v0.1) Log In

Brain Network
Linear correlation analysis of brain images to detect brain activity.
Author: Maxeler London

Correlation
Correlation is a statistical measure that indicates the extent to which two or more variables fluctuate together.
Author: Maxeler Analytics

High Speed Packet Capture
Provides sustained line-rate packet capture in distributed write mode and at bursts of up to 24GB/s in size in local.
Author: Maxeler Networking

Breast Mammogram ROI Extraction
This App extracts the region of interest from breast mammogram images. Basically, this app removes background.
Author: Faculty of Engineering, University of...

Linear Regression
In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable.
Author: Maxeler Team

Classification
Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called
Author: Maxeler Networking

Reverse Time Migration
Real time seismic monitoring of hydraulic fracturing sites, more efficient subsurface exploration and prediction.
Author: Maxeler London

Smith Waterman Demo
Smith Waterman is a standard textbook algorithm for local gene sequence alignment. While it is impractical to
Author: Maxeler London

Low-Latency HTTP Web-Server
This App implements an HTTP Web-Server in a DFE. The App serves static webpages directly from LMS.
Author: Maxeler Bangalore

N-Body Simulation
The N-Body App simulates interactions between N particles under gravitational forces in space. A particle's state
Author: Maxeler London

Dense Matrix Multiplication
A matrix is dense if most of its elements are non-zero. Multiplication involves a dot product between every row of
Author: Maxeler Analytics

Fractal
Generate the Mandelbrot and Julia sets.
Author: Maxeler London

Fast Fourier Transform 1D
This application performs a one dimensional Fast Fourier Transform on a one dimensional array of complex numbers.
Author: Maxeler London

Fast Fourier Transform 2D
This application performs a 2D Fast Fourier Transform on a two dimensional array of complex numbers.
Author: Maxeler London

Motion Estimation
Motion Estimation is used in video encoding to describe a video frame by motion vectors from other frames.
Author: Maxeler Team

Hybrid Coin Miner
A coin miner application that can mine SHA-256 based coins and Script based coins simultaneously.
Author: Maxeler London

Hessian option price
Monte Carlo options price.
Author: Maxeler Analytics

Regex
The Regex app takes a regular expression and builds a state machine to implement the regular expression.
Author: Maxeler Networking

Jacobi Solver
The Jacobi App implements a solver for equations of the type Ax=b, where A is constant but where we have a set of
Author: Maxeler London

Non-Central Chi Square
Non-Central Chi Square Distribution, that generate lots of sample from complex distribution.
Author: Maxeler California