Virtual Screening on FPGA: Performance and Energy versus Effort

Tom Vander Aa, Tom Haber, Thomas J. Ashby, Roel Wuyts, Wilfried Verachtert

ExaScience Life Lab, imec, Belgium

SC22 Dallas, TX | hpc accelerates.

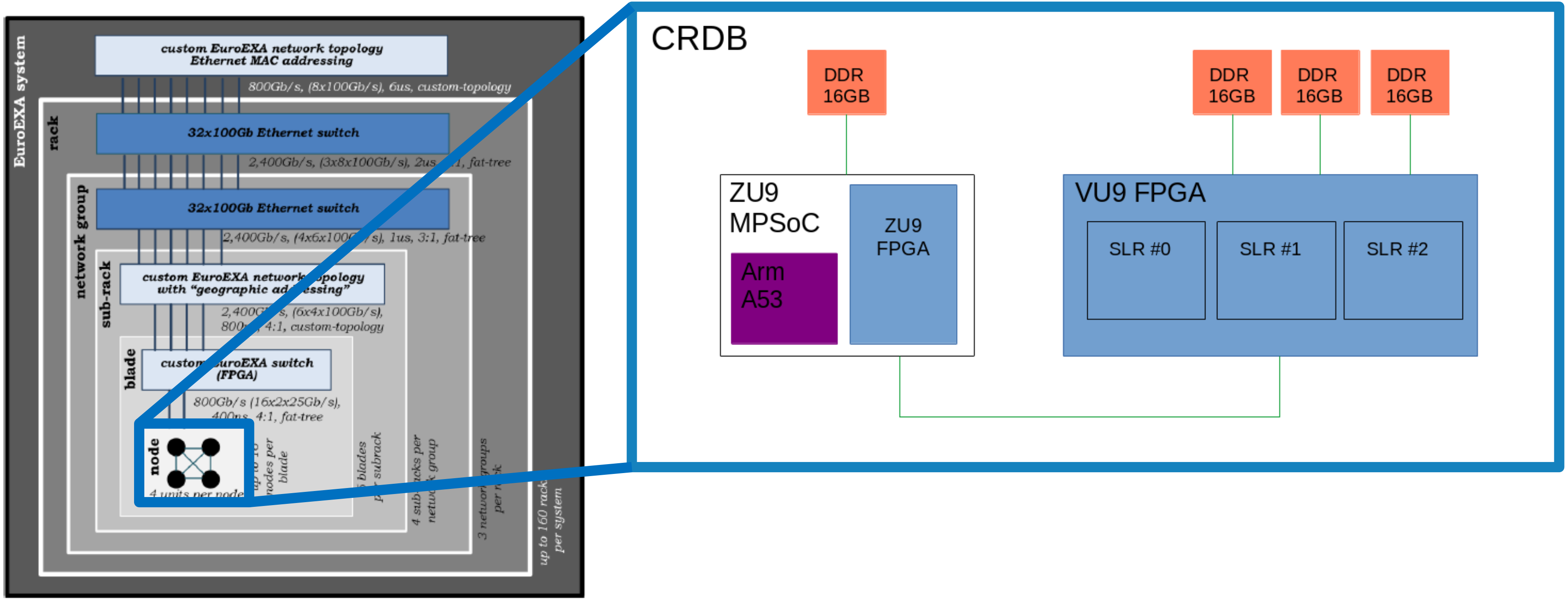
EUROEXA PROJECT
BUILDING AN FPGA-BASED SUPERCOMPUTER

- EU-funded project
  - September 2017 – December 2021 (4 years)
  - Budget €20 M
- Innovation in
  - Full system design
  - Programming Models
- Evaluation using actual HPC applications

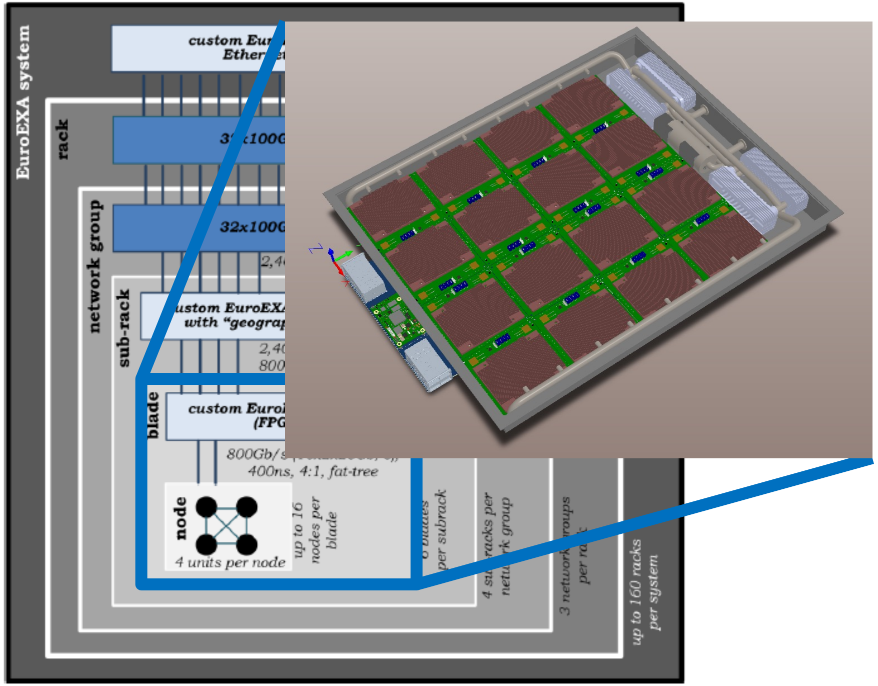
Commercial Partners

Academic/Gov. Partners

SYSTEM ARCHITECTURE AND TECHNOLOGY: COMPUTE NODE



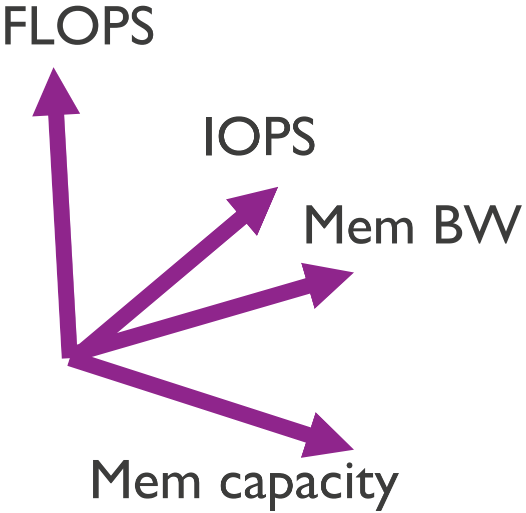
SYSTEM ARCHITECTURE AND TECHNOLOGY: BLADES

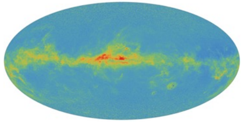


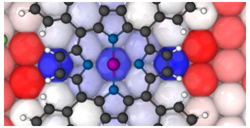
Liquid-cooled blades

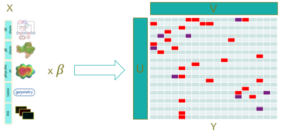
- 16 Node half depth 1u chassis
- Total Liquid Cooling technology
- 48V DC distribution
- Hot water out, chiller-less operation

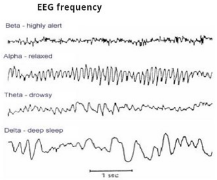
EUROEXA: CO-DESIGN, DEMONSTRATION AND EVALUATION USING EXASCALE-CLASS APPS



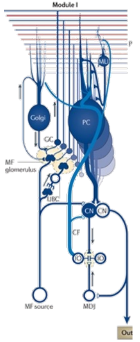

AVU-GSR


Quantum Espresso

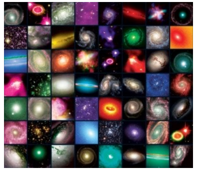

SMURFF

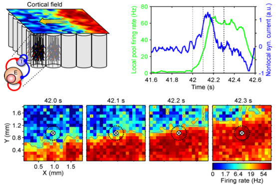


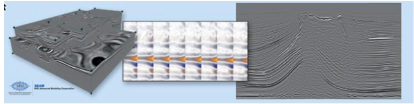
Neuromarketing

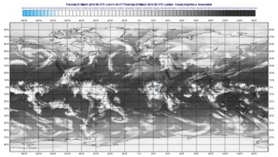

InfOli

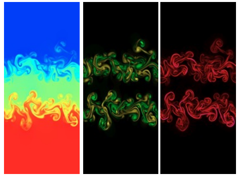

NEMO

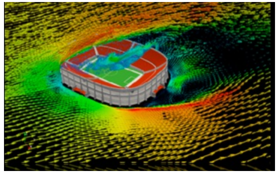

Astronomy image classification


NEST/DPSNN

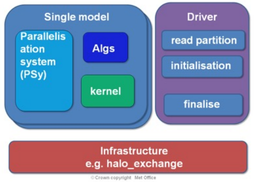

FRTM


IFS


LBM


Alya


GADGET



LFRic

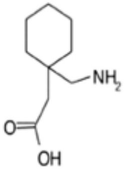
COMPOUND ACTIVITY PREDICTION

LIKE RECOMMENDER SYSTEMS

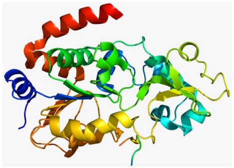
- Predict
  - compound activity on
  - protein target
  - aka chemogenomics

- Similar to
  - Netflix: users rating movies
  - Amazon: users rating books

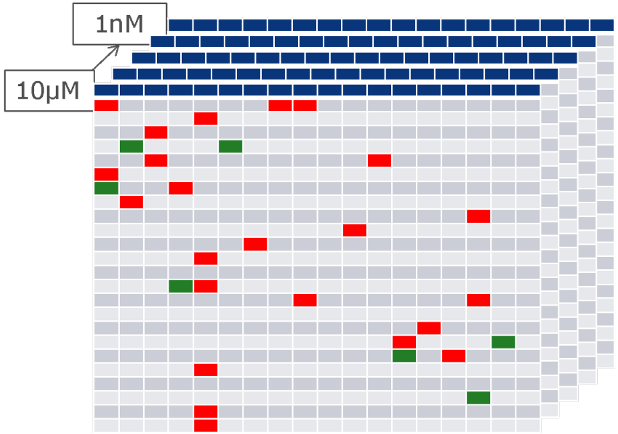
NETFLIX amazon



+



= ?

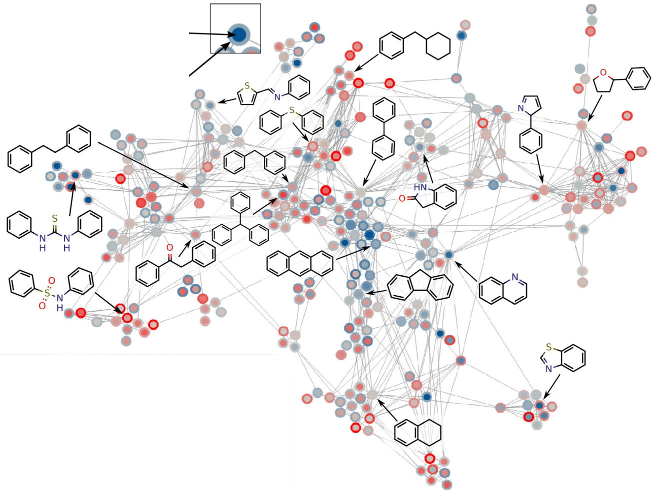


VIRTUAL MOLECULE SCREENING IS THE INFERENCE STAGE

ML PROBLEM NEEDING MASSIVE THROUGHPUT

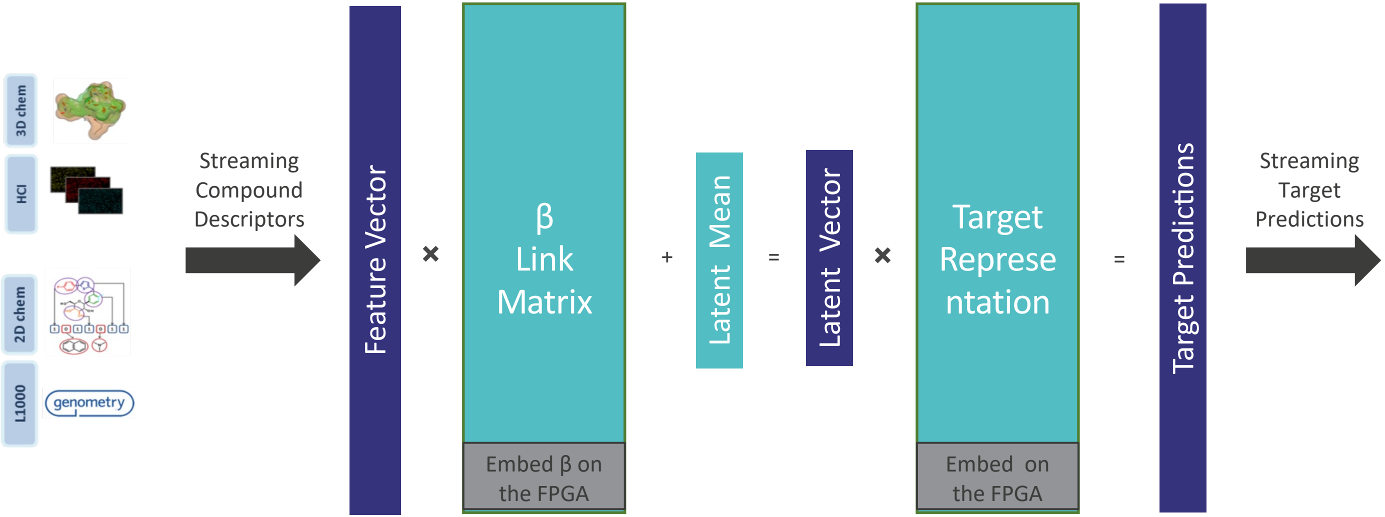
- Early stage drug discovery example
  1. Build chemogenomics model
  2. Scan space of possible chemicals for very active molecules
  3. Pass promising candidates along for investigation

- **Virtual Molecule Screening**
  - Virtual chemical space is **essentially unlimited:** 10^{60}
  - Want to scan as much as possible
  - Fast and low energy compute



VIRTUAL MOLECULE SCREENING STRUCTURE

SIMPLE LINEAR ALGEBRA PIPELINE

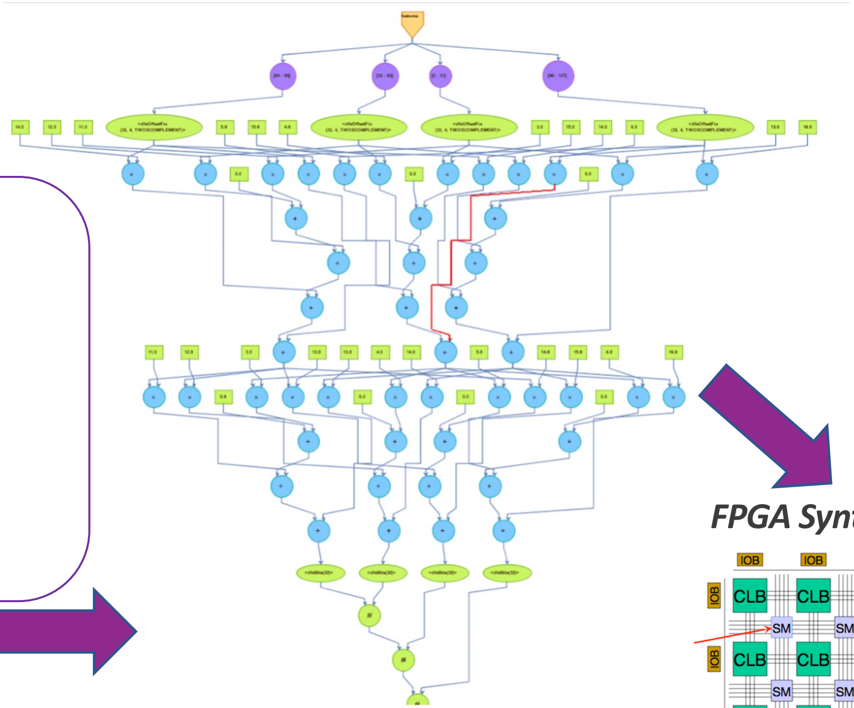


HIGH-LEVEL SYNTHESIS PLAYS AN IMPORTANT ROLW

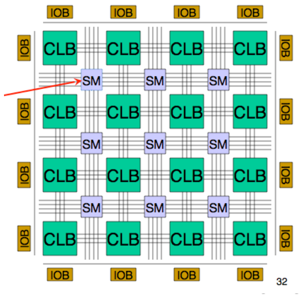
Source Code

High Level Synthesis

Convert control code
Extract Parallelism
Static Scheduling
Distribute Arrays



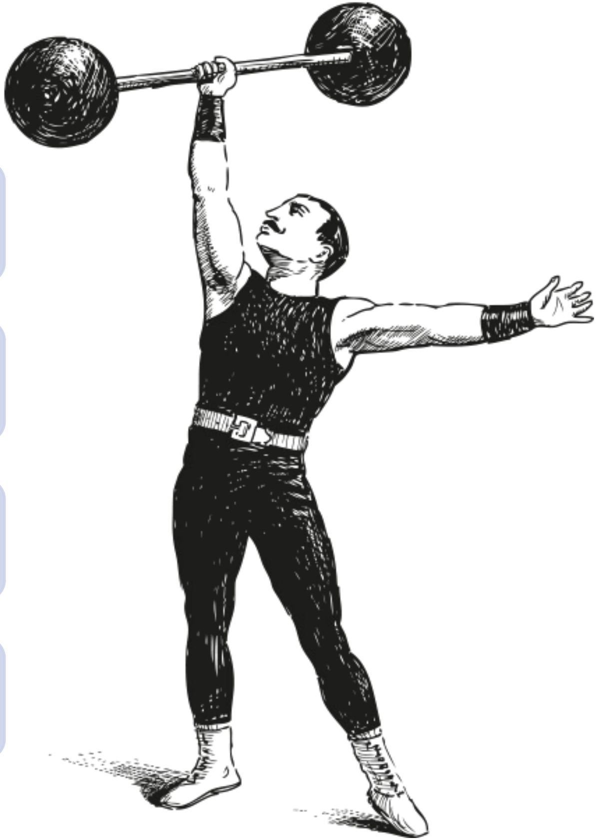
FPGA Synthesis



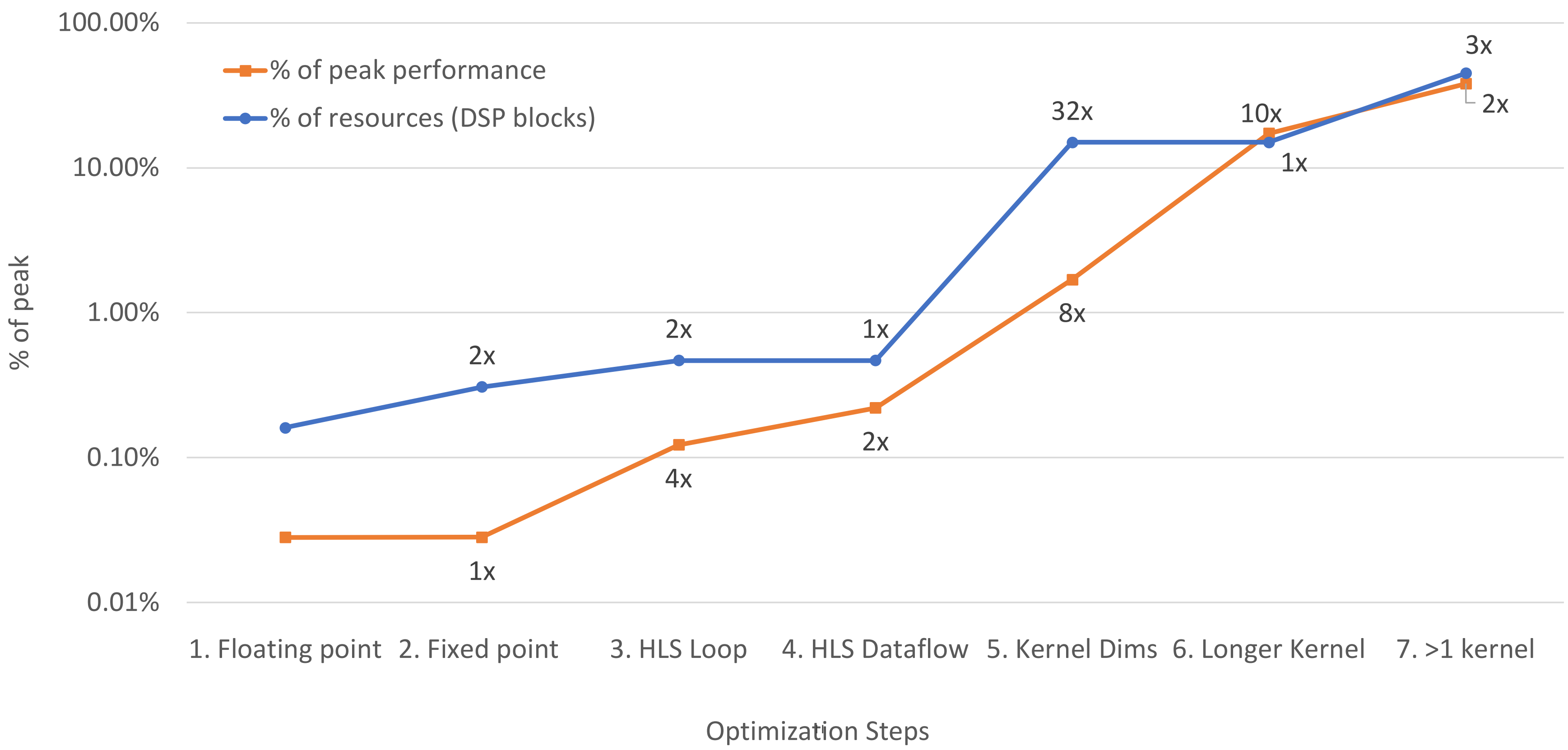
HIGH EFFORT NEEDED FOR FPGA MAPPING

WE NEED TO HELP THE HLS COMPILER

Increase Parallelism	• Inner loops are completely unrolled • Local arrays spread on FPGA
Reduce Complexity	• Trim to <100 lines of code • No branches are left
Use Local Memory	• Store model on the FPGA
Reduce Bit-Width	• 16 bit fixed point



TRANSFORMATIONS GIVE 1000X PERFORMANCE GAIN



COMPARISON TO GPU AND CPU
PERFORMANCE, ENERGY AND EFFORT

- Platforms
  - 24 core Intel Skylake CPU
  - Nvidia A100 GPU
  - Xilinx Alveo U200 FPGA

	CPU	GPU	FPGA
Peak Performance (GF/s)	3072	19500	684
Achieved Performance (GF/s)	402	3265	260
% of Peak Performance	13%	17%	38%
Measured Power Drain (Watt)	205	200	37
Energy Efficiency (GF/s/Watt)	1.8	10	3

- Results
  - Performance (% peak): FPGA is best
  - Energy Efficiency: GPU best
  - Effort: FPGA mapping was significantly more difficult
    - Long synthesis times, and timing or routing failures
    - Many optimization steps
    - Even with a background in CGRA compilers

CONCLUSIONS
NOT A GREAT SUCCESS

- EuroEXA project set out to bring scientific computing to FPGAs
- In the end very few applications managed to make good use of FPGA
- Code transformations improve performance 1000x, with large effort
- Yet, in pure performance and energy efficiency we cannot beat GPUs
- I would not call this a success…

QUESTIONS ?

embracing a better life