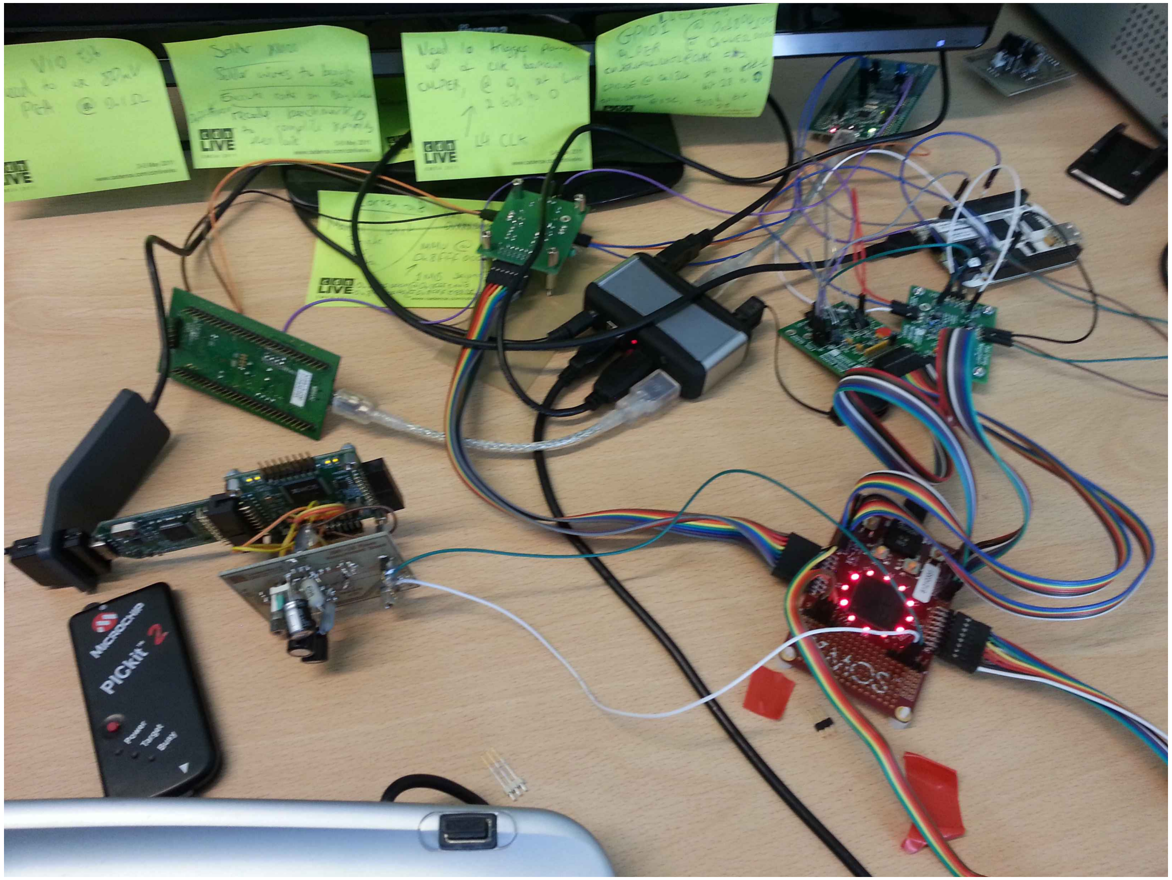
Impact of different compiler options on energy consumption



James Pallister
University of Bristol / Embecosm

Simon Hollis
University of Bristol

Jeremy Bennett
Embecosm

Motivation

- Compiler optimizations are claimed to have a large impact on software:
  - Performance
  - Energy
- No *extensive* study prior to this considering:
  - Different benchmarks
  - Many individual optimizations
  - Different platforms
- This work looks at the effect of many different optimizations across 10 benchmarks and 5 platforms.
- 238 Optimization passes covered by 150 flags
  - Huge amount of combinations

This Talk

- This talk will cover:
  - Importance of benchmarks
  - How to explore 2^150 combinations of options
  - Correlation between time and energy
  - How to predict the effect of the optimizations
  - The best optimizations

Importance of Benchmarks

- One benchmark can't trigger all optimizations
- Perform differently on different platforms
- Need a range of benchmarks

- Broad categories to be considered for a benchmark:
  - Integer
  - Floating point
  - Branching
  - Memory

Existing Benchmark Suites Considered

- **MiBench**
- **WCET**
- **DSPstone**
- **ParMiBench**
- **OpenBench**
- **LINPACK**
- **Livermore Fortran Kernels**
- **Dhry/Whet-stone**

- Require embedded Linux
- Targeted at higher-end systems
- Multithreaded benchmarks typically for HPC
- Don't necessarily test all corners of the platform

Our Benchmark List

Name	Source	B	M	I	FP	T	License	Category
Blowfish	MiBench	L	M	H	L	Multi	GPL	security
CRC32	MiBench	M	L	H	L	Single	GPL	network, telecomm
Cubic root solver	MiBench	L	M	H	L	Single	GPL	automotive
Dijkstra	MiBench	M	L	H	L	Multi	GPL	network
FDCT	WCET	H	H	L	H	Single	None†	consumer
Float Matmult	WCET	M	H	M	M	Single	GPL	automotive, consumer
Integer Matmult	WCET	M	M	H	L	Single	None†	automotive
Rjindael	MiBench	H	L	M	L	Multi	GPL	security
SHA	MiBench	H	M	M	L	Multi	GPL	network, security
2D FIR	WCET	H	M	L	H	Single	None†	automotive, consumer

Choosing the Platforms

- Range of different features in the platforms chosen
  - Pipeline Depth
  - Multi- vs Single- core
  - FPU available?
  - Caching
  - On-chip vs off-chip memory

Platforms Chosen

ARM Cortex-M0	ARM Cortex-M3	ARM Cortex-A8	XMOS L1	Adapteva Epiphany
Small memory	Small memory	Large memory	Small memory	On-chip and off-chip memory
Simple Pipeline	Simple Pipeline, with forwarding logic, etc.	Complex superscalar pipeline	Simple pipeline	Simple superscalar pipeline
		SIMD/FPU		FPU
			Multiple threads	16 cores

Experimental Methodology

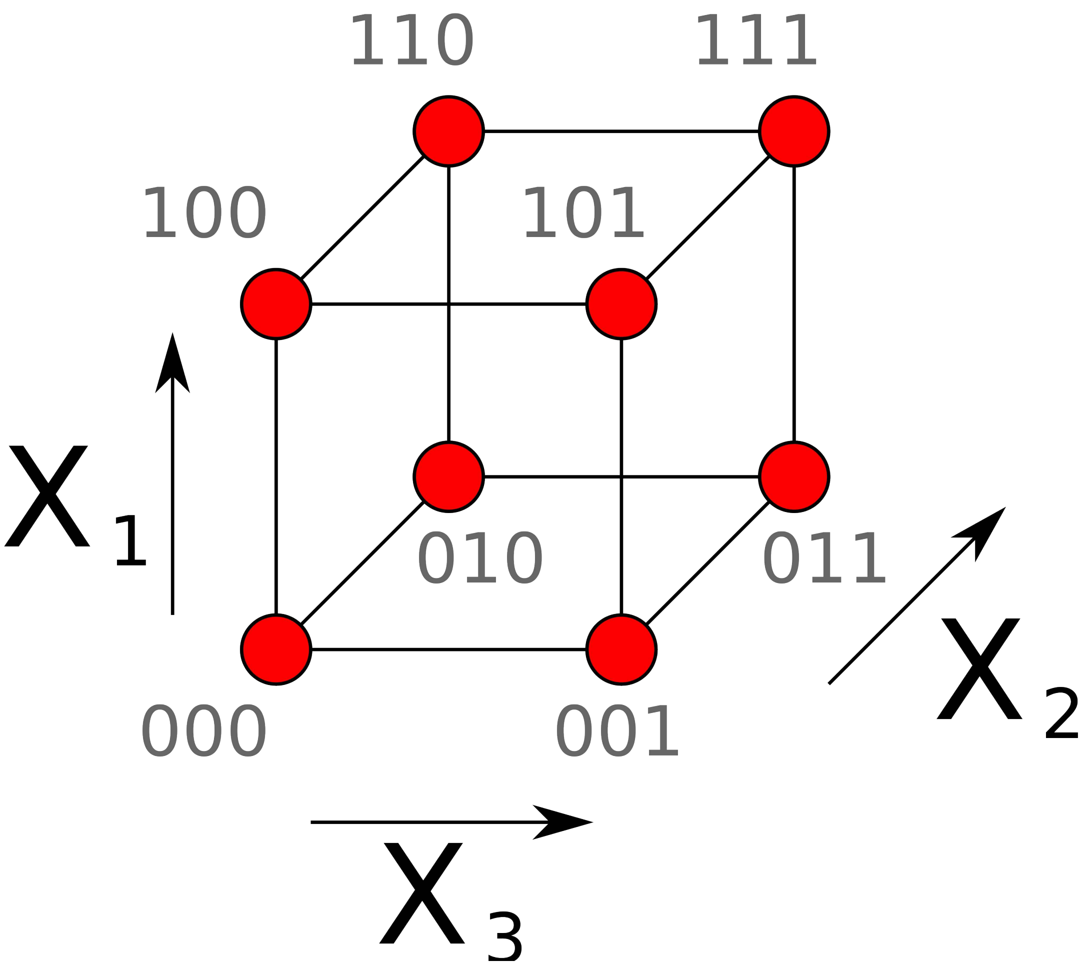
- Compiler optimizations have many non-linear interactions
- 238 optimization passes combined into 150 different options (GCC)
- 82 compiler options enabled by O3
- How to test all of these, while accounting for the interactions between optimizations?

Fractional Factorial Designs

Full Factorial Design

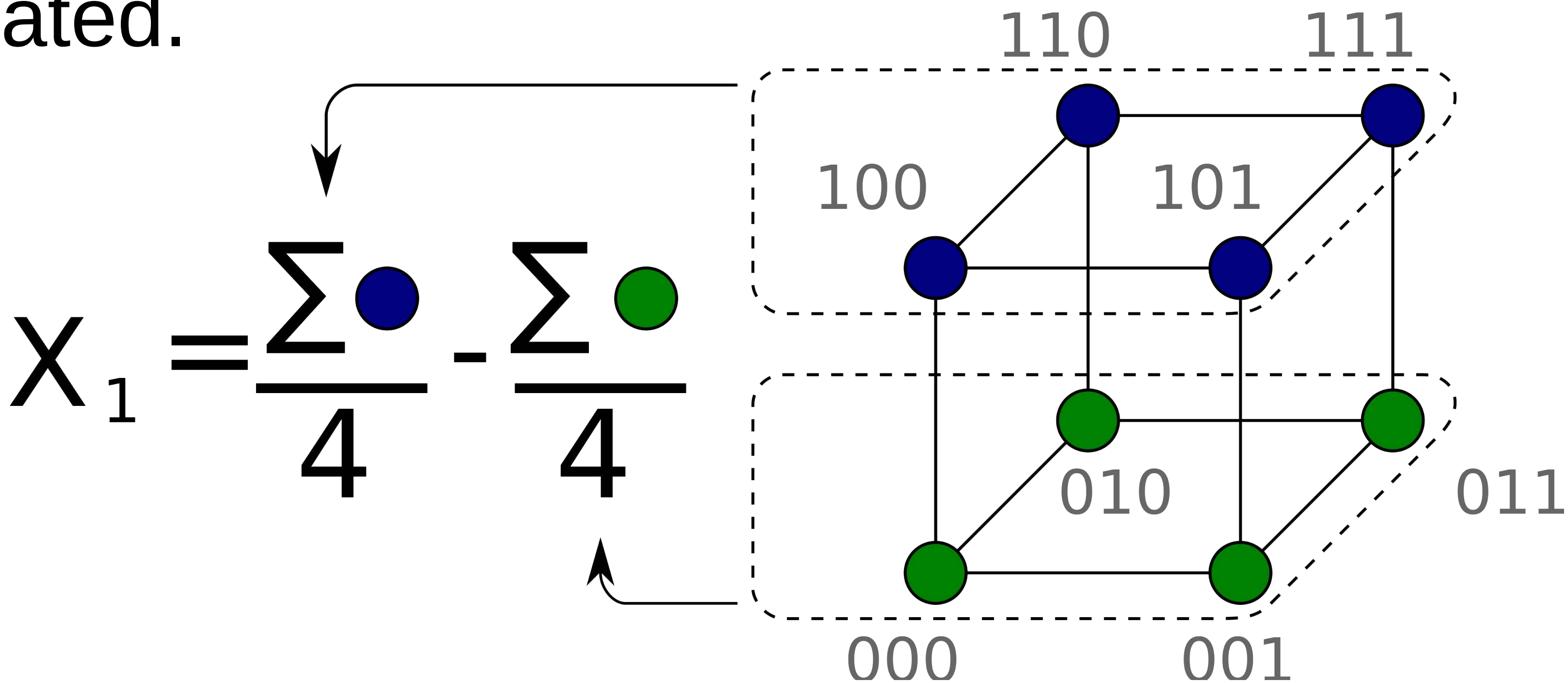
Example:

- 3 options to investigate
- Each option can be on or off (2 level)
- 2^3 tests to be run



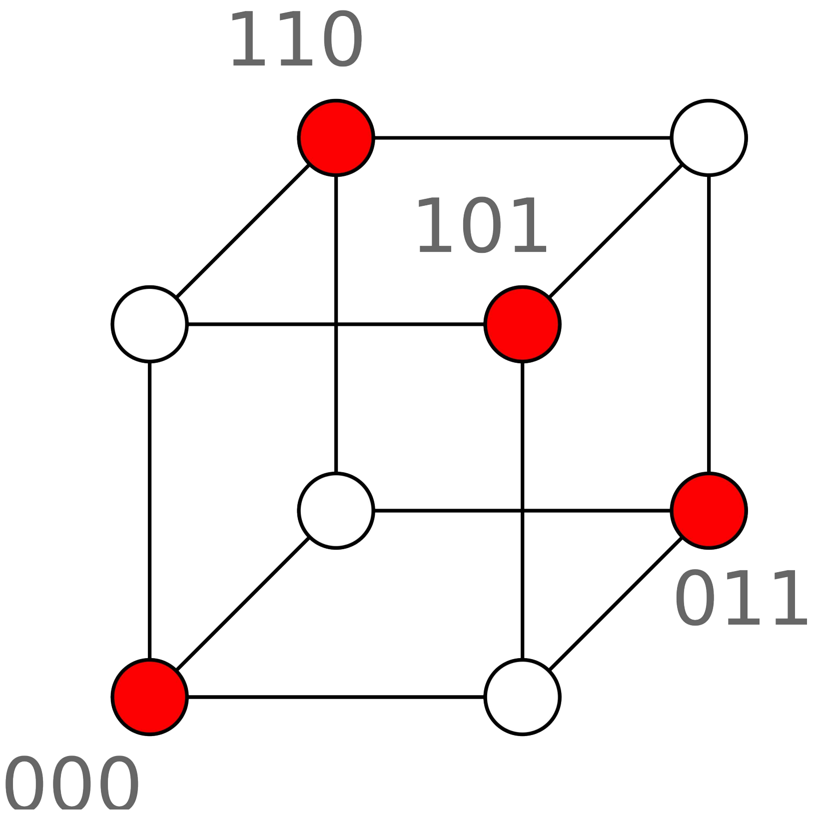
Estimating an Option's Effect

- The effect of a single option can be calculated.



Fractional Factorial Design

- Use a subset of the full factorial design
- Shown here is a 'half fraction'
- 2^(3-1) tests to be run



Loss of Information

- Less runs = less information
- The fewer runs performed, the fewer interactions can be resolved
- The 'resolution' of the fractional factorial design

O1 flags (37 factors)

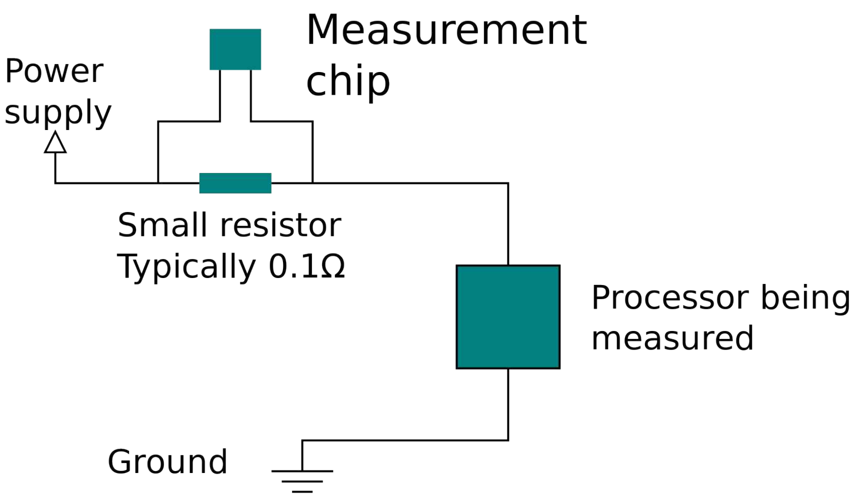
Resolution	Runs Needed
3	256
4	1024
5	2048
6	4096
Full	137438953472

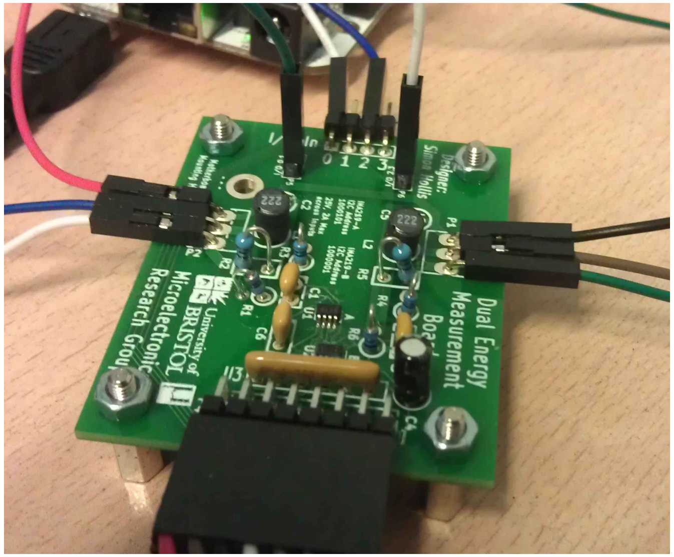
10 hours

77000 years

Hardware Measurements

- Current, voltage and power monitor
- 10 kSamples/s
- Low noise
- XMOS board to control and timestamp measurements
- Integrate to get energy consumption





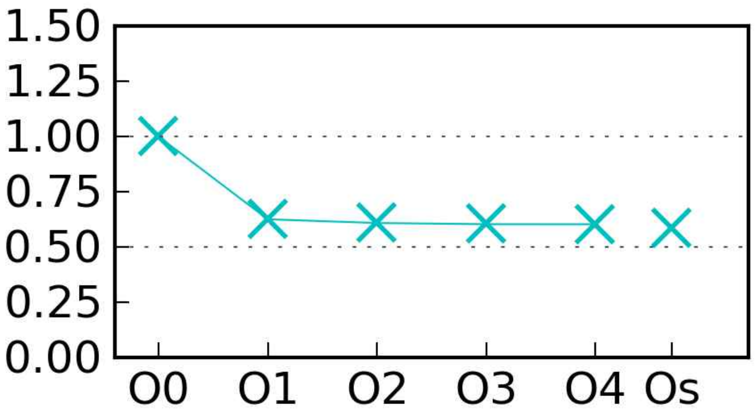
Results

- Energy consumption ≈ Execution time
  - Generalization, not true in every case
- Optimization unpredictability

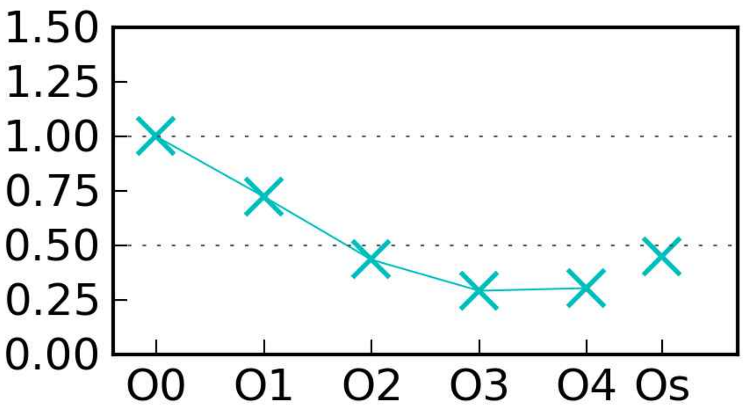
- No optimization is universally good across benchmarks and platforms

Overview

FDCT, Cortex-M0



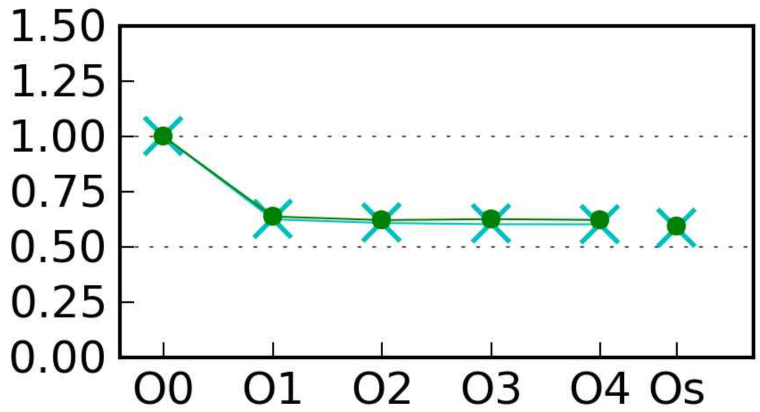
FDCT, Cortex-A8



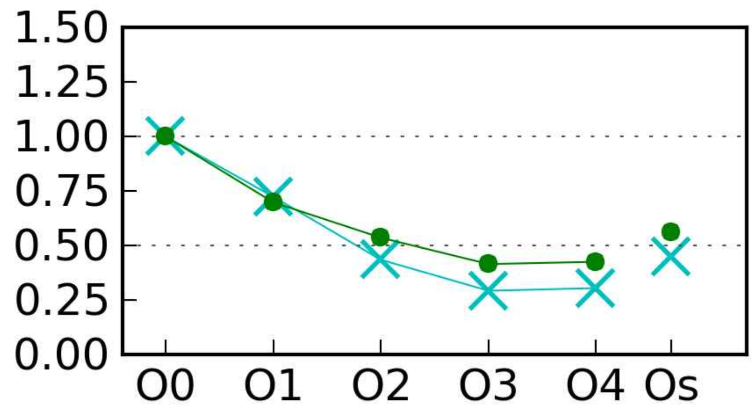
Execution time

Overview

FDCT, Cortex-M0



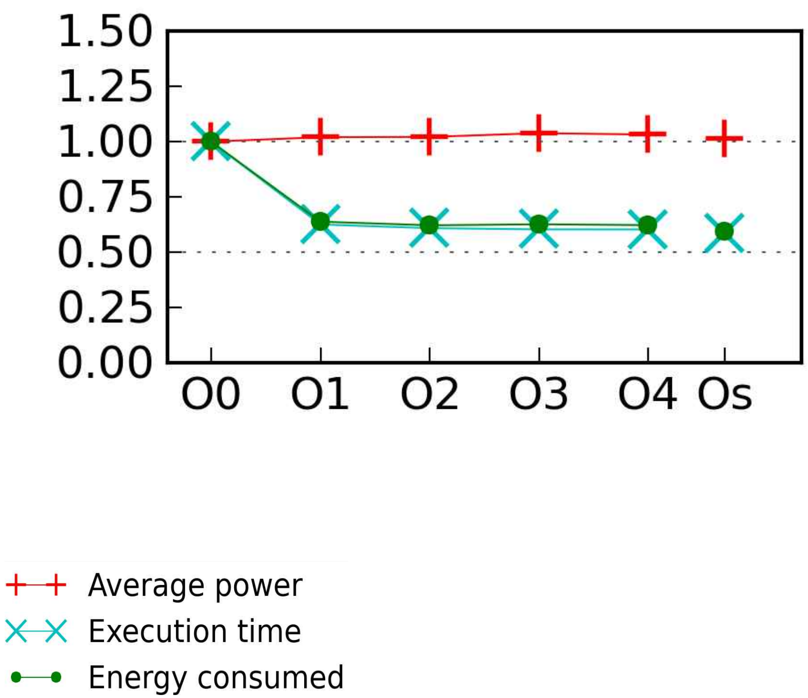
FDCT, Cortex-A8



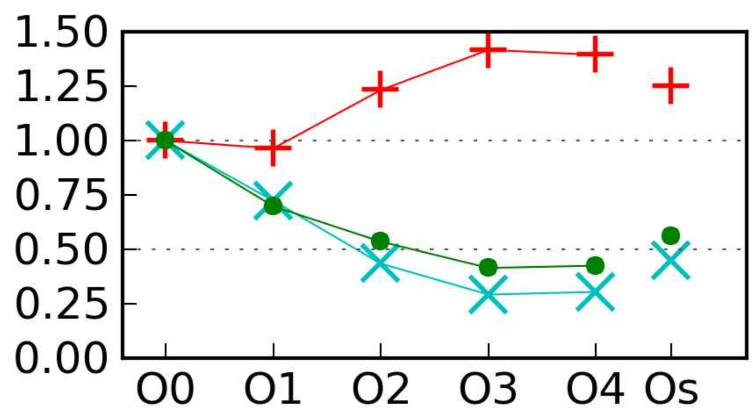
Execution time
Energy consumed

Overview

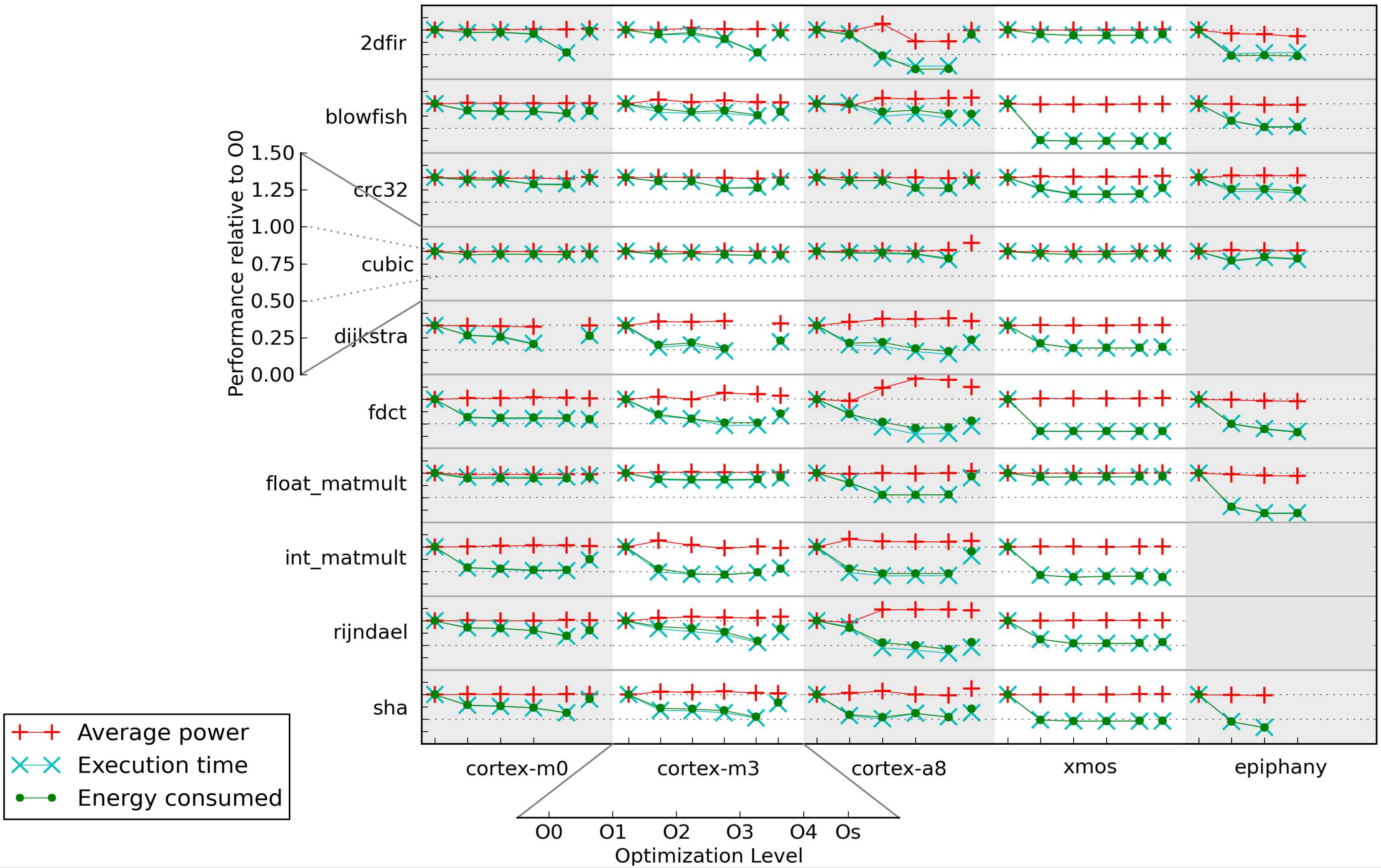
FDCT, Cortex-M0



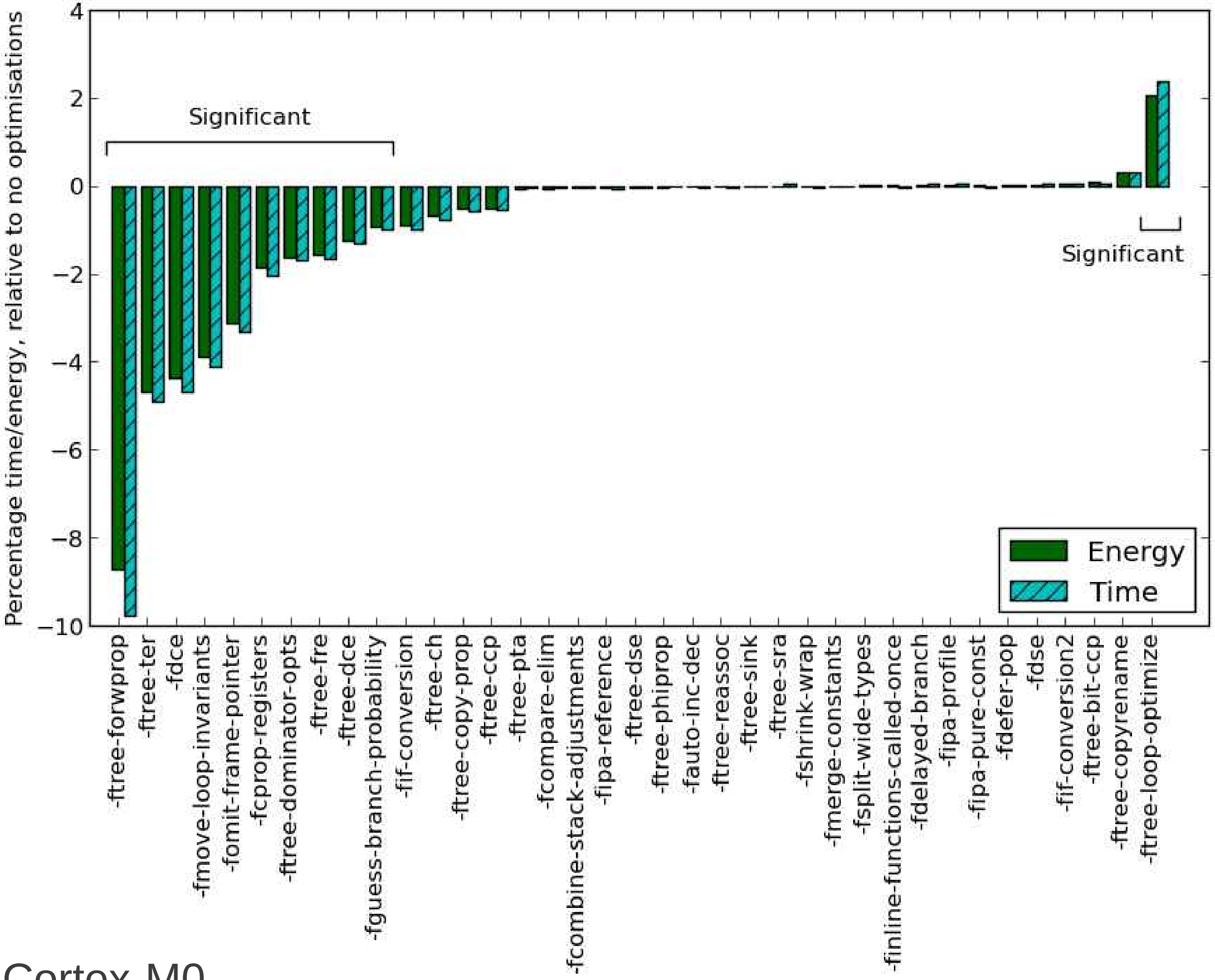
FDCT, Cortex-A8



Overview

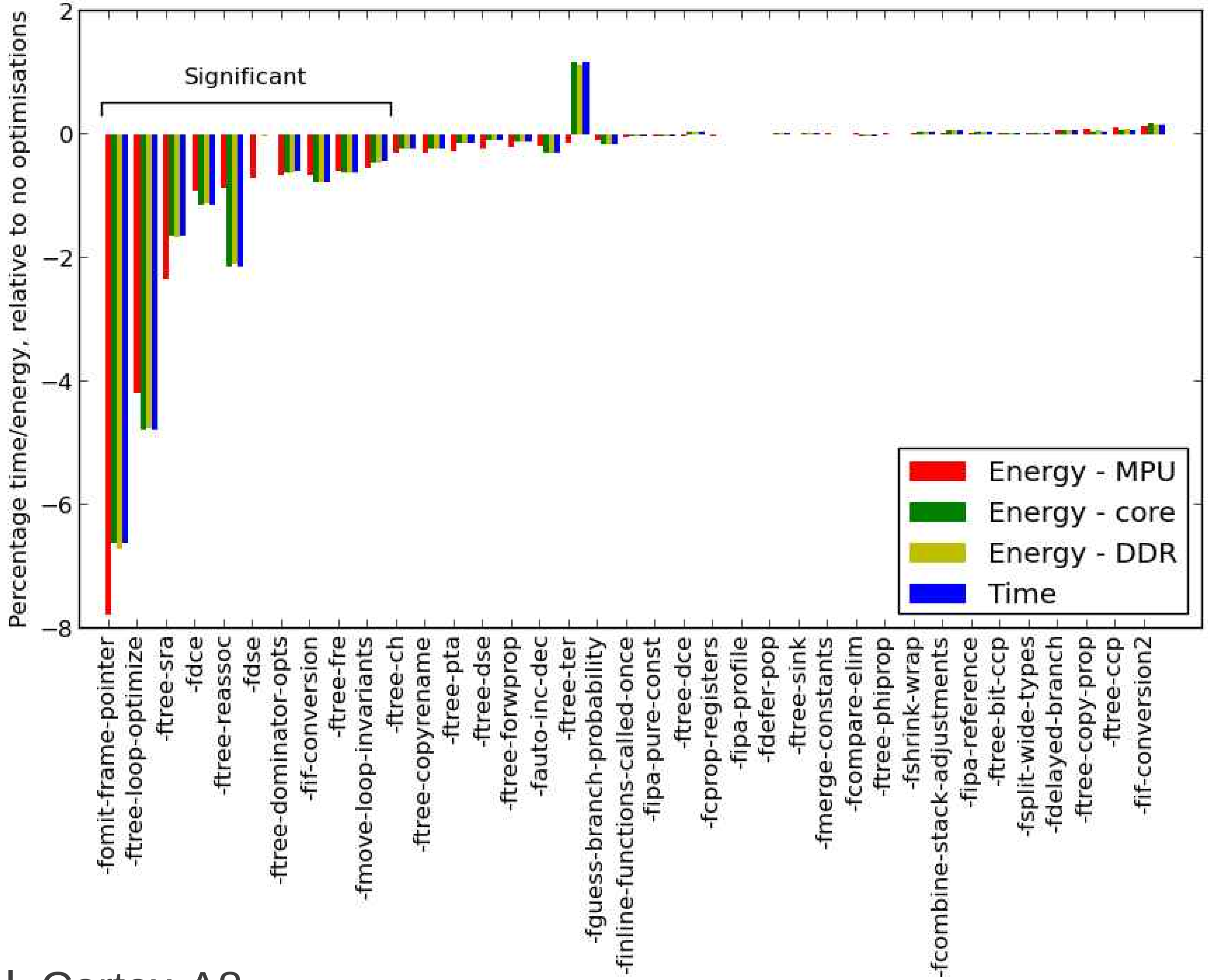

2dfir
blowfish
crc32
cubic
dijkstra
fdct
float_matmult
int_matmult
rijndael
sha
Performance relative to O0
1.50
1.25
1.00
0.75
0.50
0.25
0.00
cortex-m0
cortex-m3
cortex-a8
xmos
epiphany
O0
O1
O2
O3
O4
Os
Optimization Level
Average power
Execution time
Energy consumed

Time ≈ Energy



O1 Flags, FDCT, Cortex-M0

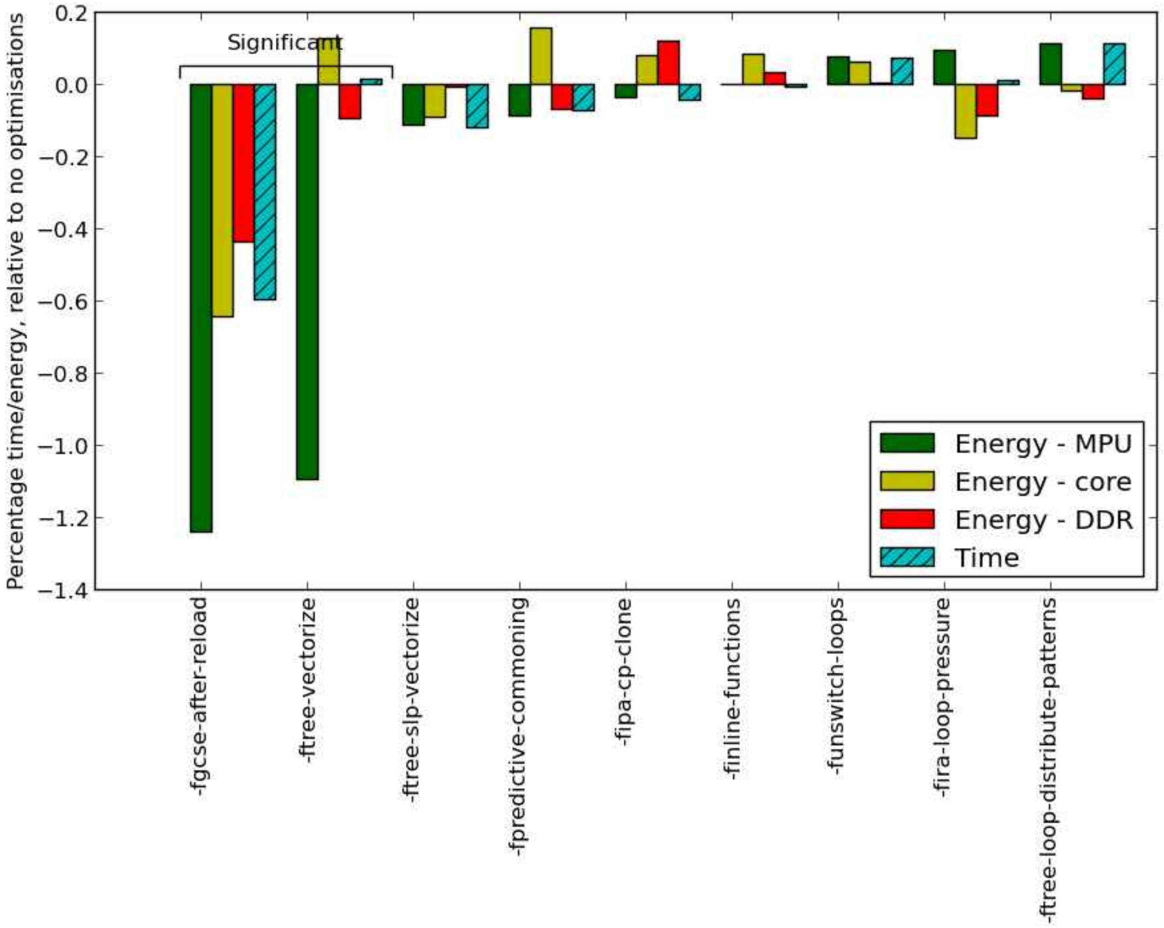
Less Correlation



O1 Flags, Rijndael, Cortex-A8

When Time ≠ Energy

- Complex pipeline
- -ftree-vectorize
  - NEON SIMD unit
  - Much lower power



O3 Flags, 2DFIR, Cortex-A8

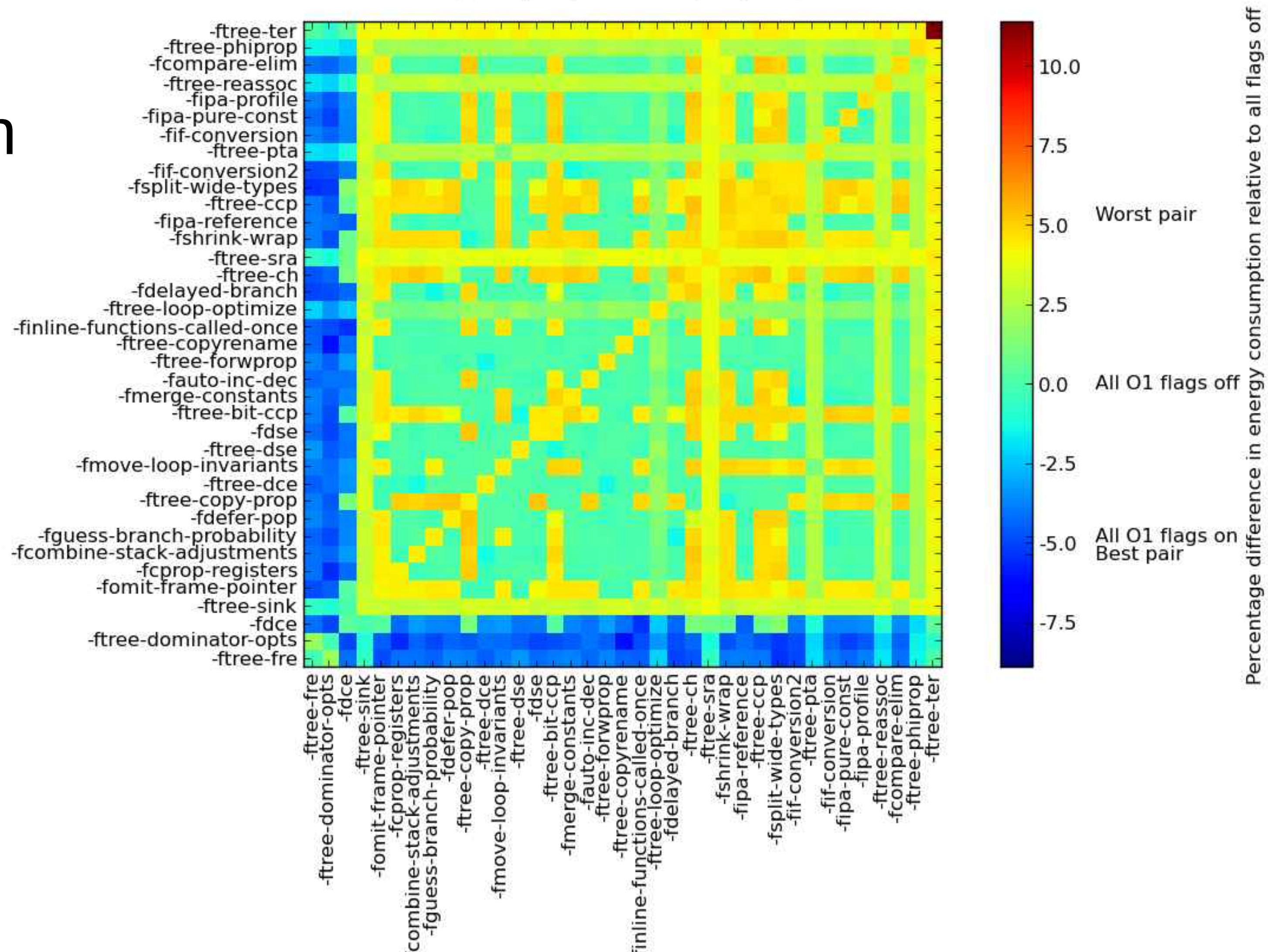
Conclusion: Mostly, Time ≈ Energy

- Highly correlated
- Especially so for 'simple' pipelines
- Little scope for stalling or superscalar execution

- Complex pipelines:
  - Still a correlation
  - But more variability
  - SIMD, superscalar execution
- To get the most optimal energy consumption we need better than "go fast"

Optimization Unpredictability

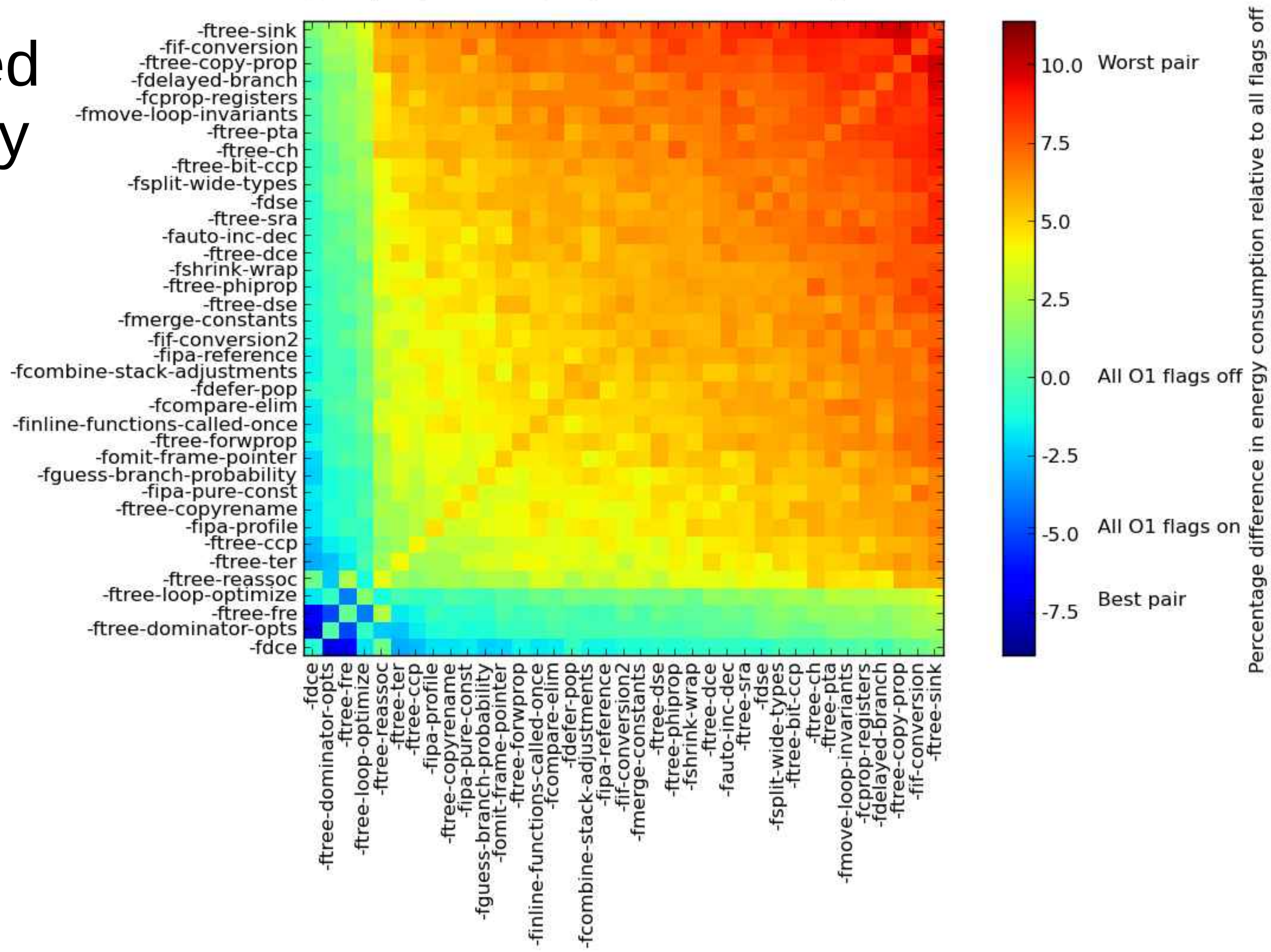
- Pairs of optimizations on top of O0
- Possibly higher order interactions occurring?



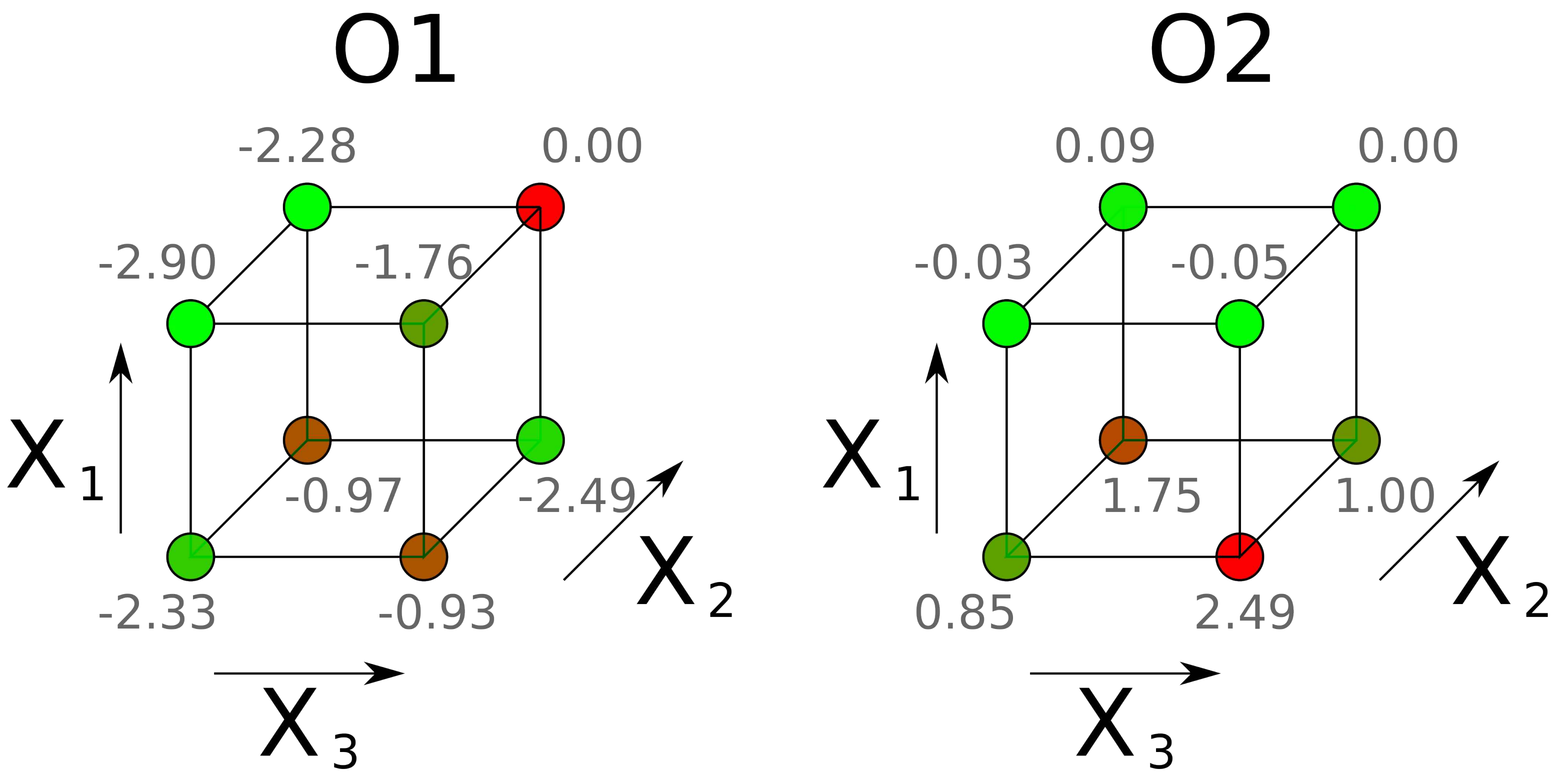
O1 Flags, Cubic, Cortex-M0

Modelled

- Model constructed from 1 and 2 -way interactions
- Doesn't predict very well



Case Study: Interactions



X_1 -fguess-branch-probability

X_2 -ftree-dominator-opts

X_3 -ftree-ch

The Best Three Optimizations for Energy

Benchmark	Cortex-M0	Cortex-M3	Cortex-A8	Epiphany
2dfir	E	T, G, H	N, G, C	H, A, D
blowfish	B, J, E	J, B, G	K, B, E	D, P, H
crc32	F	F	F, G	
cubic	A, I	A, I	A	A, I, O
dijkstra	I, A, B	F, I, A	F, I, A	
fdct	J, G, D	J, G, K	M, K, J	A, H, D
float_matmult	C, E	C, E, G	N, L	D, H, A
int_matmult	C, E, B	C, L, F	L, N, M	A, H, D
rijndael		B, C, R	K, B, S	
sha	B, C, E	C, B, F	C, B, M	D, C, Q

ID	Count	Flag	ID	Count	Flag	ID	Count	Flag
A	11	-ftree-dominator-opts	B	10	-fomit-frame-pointer	C	10	-ftree-loop-optimize
D	7	-fdce	E	7	-fguess-branch-probability	F	7	-fmove-loop-invariants
G	7	-ftree-ter	H	6	-ftree-ch	I	6	-ftree-fre
J	5	-ftree-forwprop	K	4	-fschedule-insns	L	3	-finline-small-functions
M	3	-fschedule-insns2	N	3	-ftree-pre	O	1	-fcombine-stack-adjustments
P	1	-fipa-profile	Q	1	-ftree-pta	R	1	-ftree-sra
S	1	-fgcse	T	1	-fpeephole2			

Conclusion: Which optimization to choose?

For the general case, this question can't be answered

- Unpredictable interactions
- Many non-linear effects
- Not enough data recorded in the fractional factorial design to model

- Evidence of higher order interactions between optimizations?

University of BRISTOL
EMBECOSM®

Conclusion: Optimizations are common across architectures...

… Sometimes

- Common options across all the ARM platforms for a particular benchmark
- A few consistently good options for Epiphany
  - Simpler instruction set
  - Newer compiler
  - Many more registers than ARM

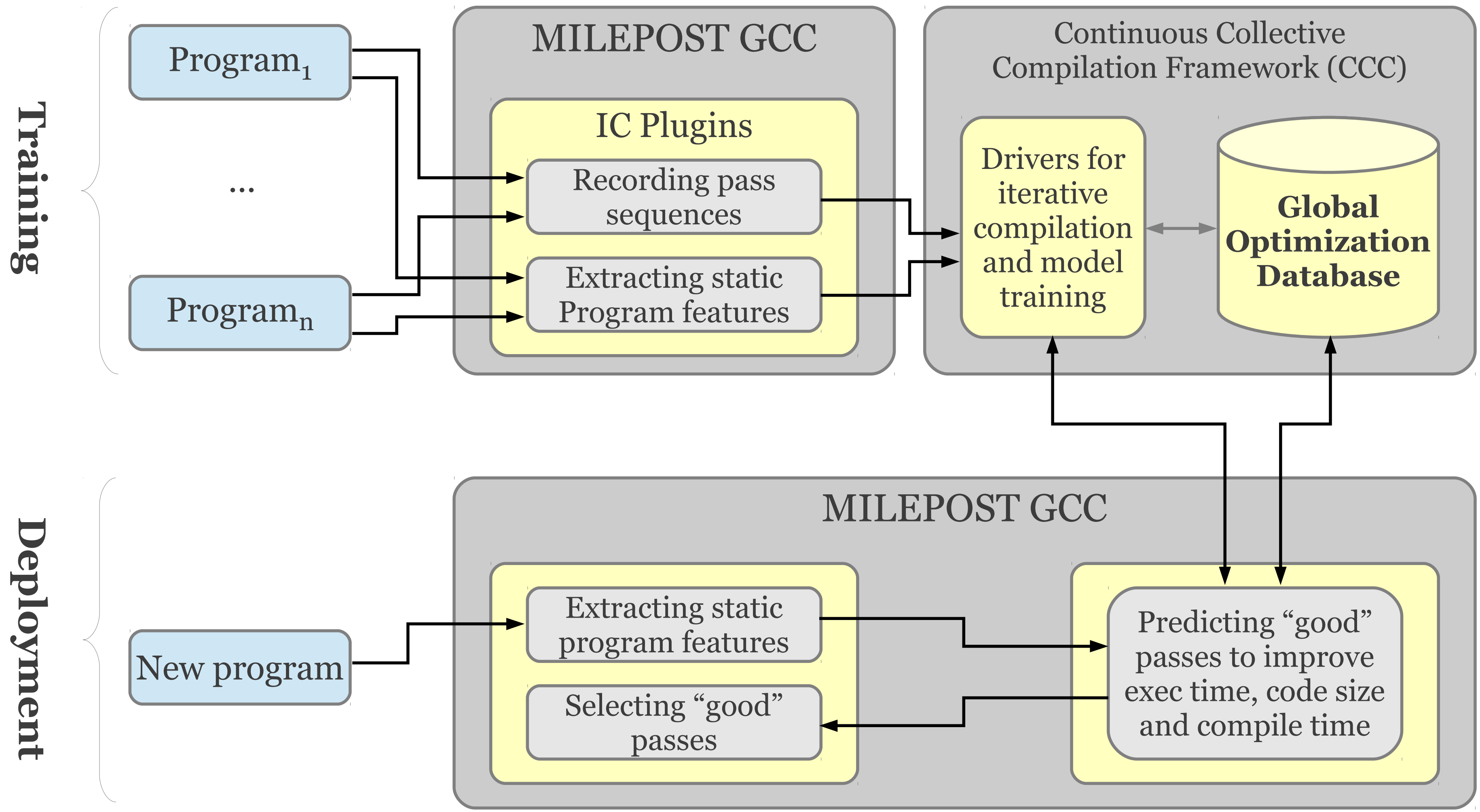
What does this mean?

For the Compiler Writer

- Current optimization levels (O1, O2, etc.) are a good balance between compile time and performance/energy.
- Never completely optimal
- Machine learning
  - MILEPOST
  - Genetic algorithms

- Current optimizations targeted for performances
- Few (if any) optimizations in current compilers designed to reduce energy consumption

MILEPOST GCC



From Fursin et al, 2008

Conclusion

- Time ≈ Energy
  - True for simple pipelines
  - Mostly true for complex pipelines
  - Good approximation
- Optimization unpredictability
  - Difficult to model the interactions between optimizations
- Commonality across platforms
  - Instruction set plays a role
  - Common options for the ARM platforms, but not Epiphany

Questions?

james.pallister@bristol.ac.uk

simon@cs.bris.ac.uk

jeremy.bennett@embecosm.com

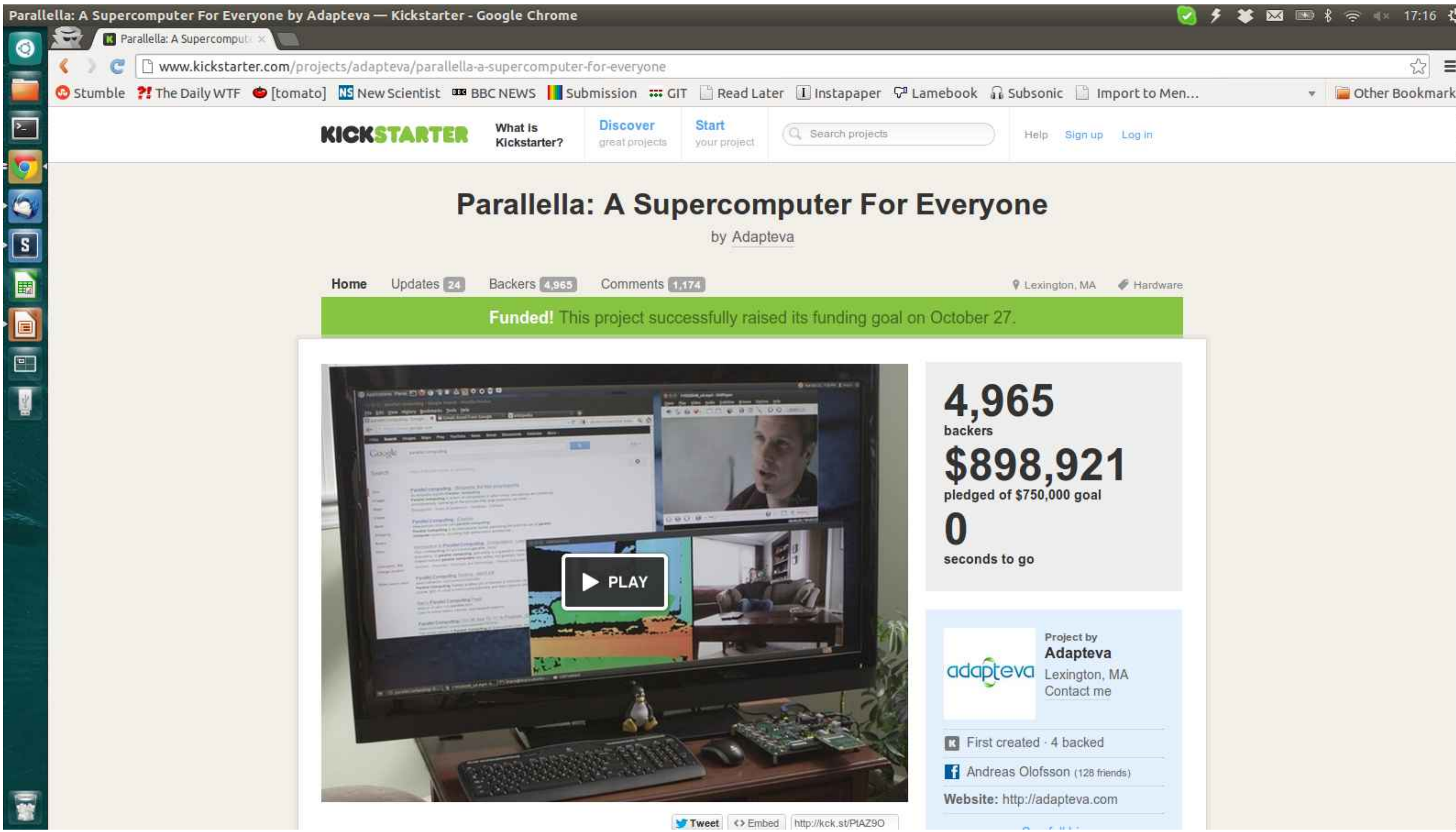
All data at: www.jpallister.com/wiki

Howto: Funding Research at the University of Bristol

Jeremy Bennett, Embecosm

Slides for NMI, 8th November 2012

Parallella



University of Bristol

EMBECOSM®

Project Organization and Funding

- A ***fully open*** research project
  - all the programs & results available as open source for download
  - all papers will be published in open access journals (£2k each)
- Funded directly by Embecosm (approx £12k)
  - paid for staff (at commercial rates as employees)
  - paid for open access publication and some equipment
- Supported by Bristol University
  - provided laboratory space and most equipment
  - provided academic supervision (Dr Simon Hollis)
- Supported by industry
  - Epiphany board (value $US 10k) loaned by Adapteva Inc.
- Supported by government
  - 27.5% R&D Tax Credit

Why Fund This Way

- Simple to set up and run
  - agreement by email
  - fortnightly progress meetings
- Fast
  - concept proposed in April 2012, started project < 3 months later
- Flexible
  - no problem using Embecosm staff at commercial rates
- Cost effective (at least for a small project)
  - no collaboration contract, no reporting bureaucracy
  - 27.5% R&D Tax Credit (more for big companies, even Starbucks)

Future Funding

- Technology Strategy Board (TSB)
  - government innovation agency
  - energy efficient computing (EEC) funding call
    - £1.25M, up to approx £150k costs per project
    - business led, consortia of 2 or more
  - 100% funding of Universities, up to 75% funding of businesses
    - *plus* R&D tax credit on top
- Joint proposal from Embecosm and Bristol University
  - develop MILEPOST concept for energy
    - but less integrated to individual compilers, use GCC and LLVM
  - write compiler passes specifically for energy saving
    - existing passes focus on code speed and size
  - measure on a range of hardware
    - does it work?