Energy-Optimal Computing in the AI Age: *Where We Are, Where Can We Go*

Christos Lamprakos
PhD Student
cplamprakos@microlab.ntua.gr

Better: **Software Execution**

Energy-Optimal ~~Computing~~ in the AI Age:
Where We Are, Where Can We Go

Christos Lamprakos
PhD Student
cplamprakos@microlab.ntua.gr

Why bother?

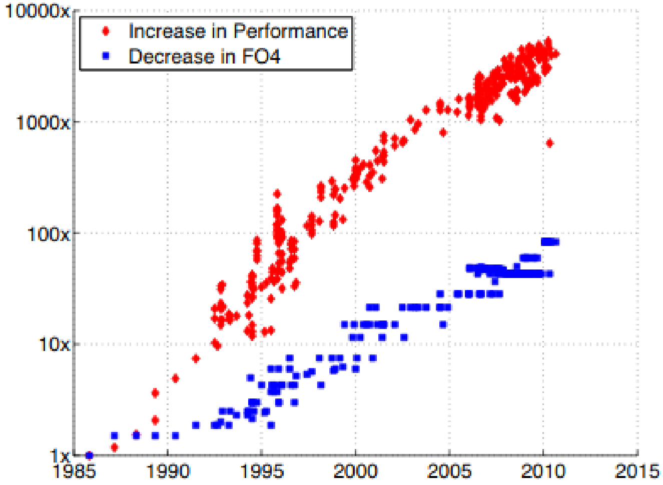


Figure 1.1.1: Improvement in microprocessor and gate performance vs. year.

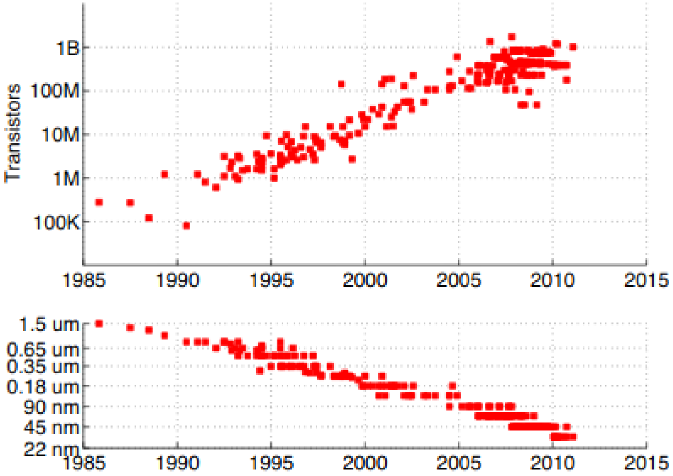


Figure 1.1.2: Number of transistors and feature size vs. year.

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

Why bother?

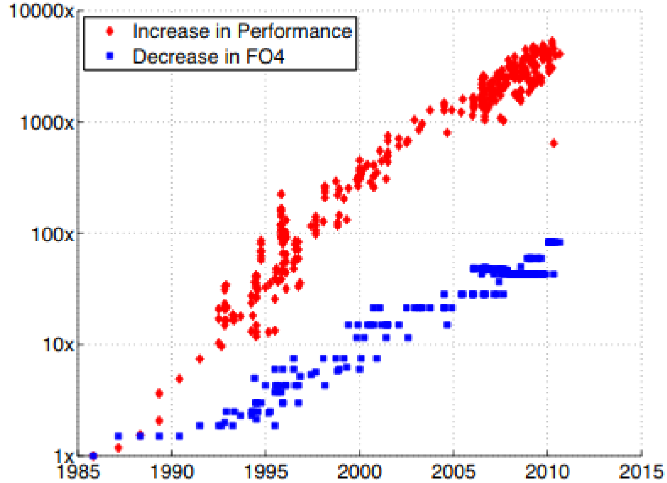


Figure 1.1.1: Improvement in microprocessor and gate performance vs. year.

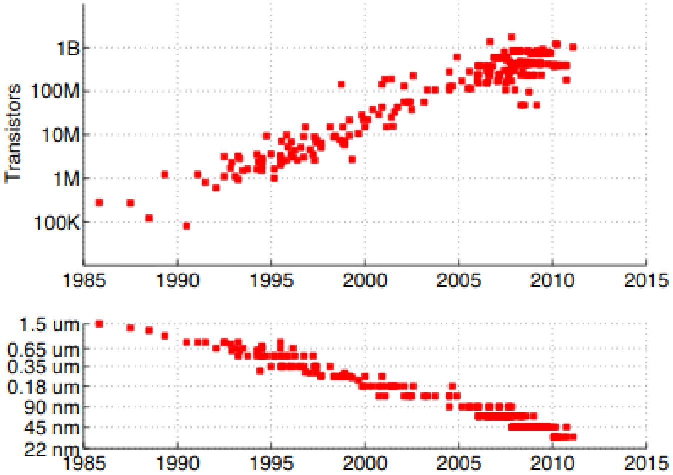


Figure 1.1.2: Number of transistors and feature size vs. year.

Performance has scaled well . . .

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

Why bother?

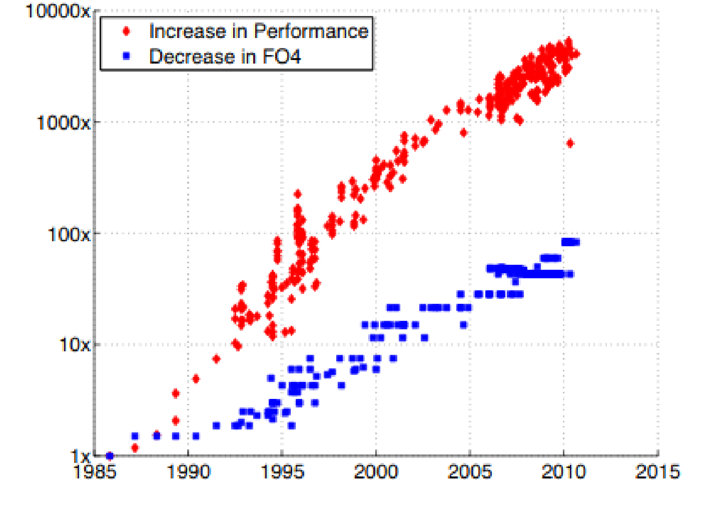


Figure 1.1.1: Improvement in microprocessor and gate performance vs. year.

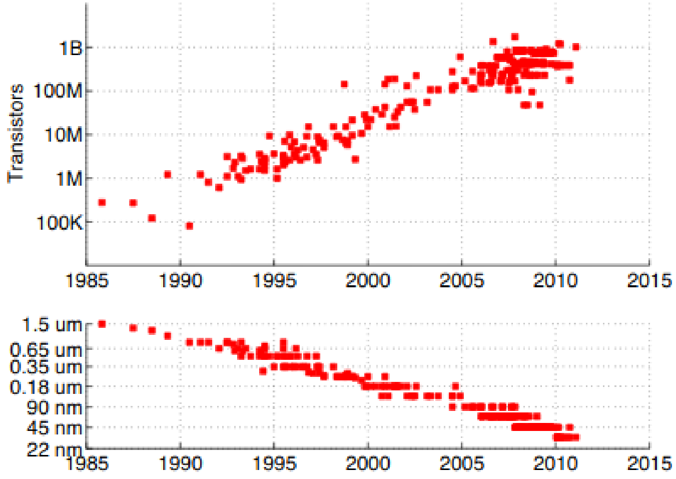


Figure 1.1.2: Number of transistors and feature size vs. year.

Transistors got
more + smaller . . .

Performance has scaled well . . .

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

. . . BUT:

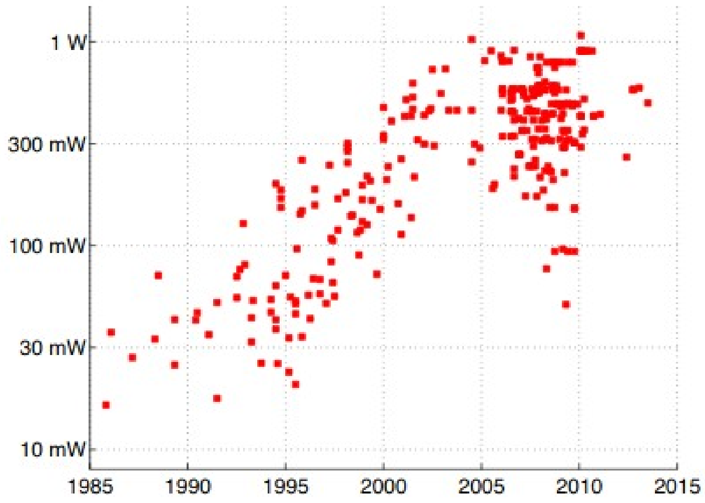


Figure 1.1.3: Power density in mW/mm² vs. year.

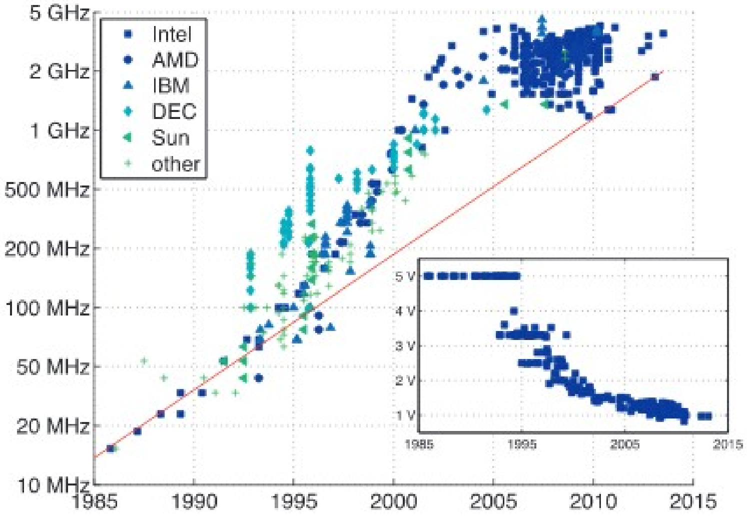


Figure 1.1.4: Clock frequency vs. year. The red line indicates frequency increase due to gate speed. The insert plot is Vdd vs. year.

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

. . . BUT:

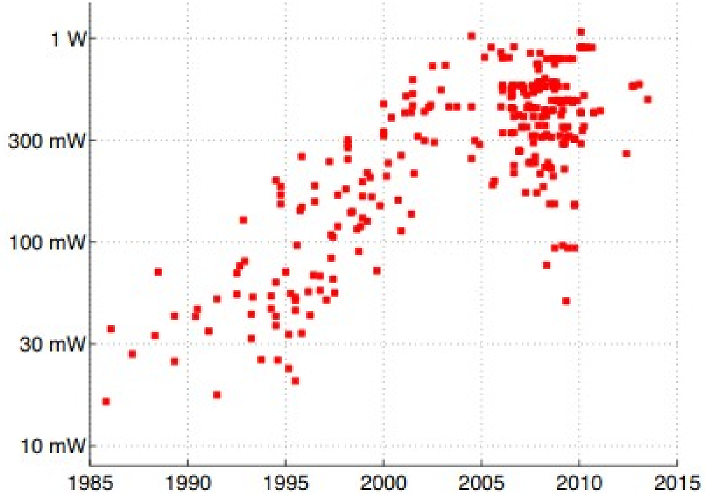


Figure 1.1.3: Power density in mW/mm² vs. year.

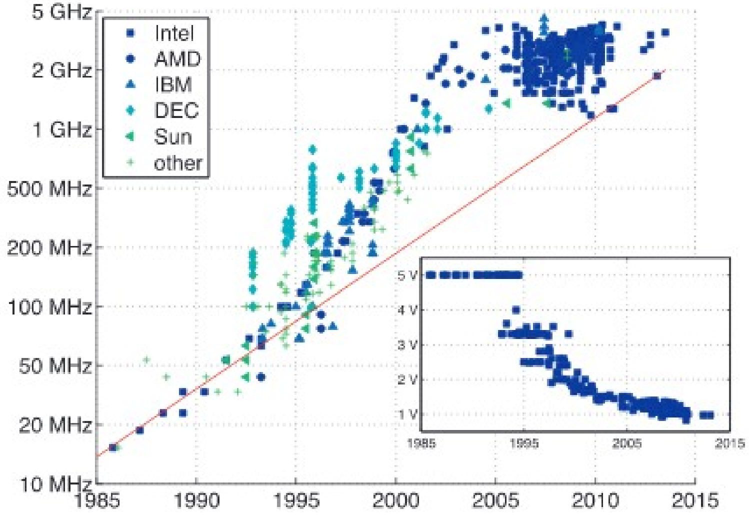


Figure 1.1.4: Clock frequency vs. year. The red line indicates frequency increase due to gate speed. The insert plot is Vdd vs. year.

We hit the power wall!

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

. . . BUT:

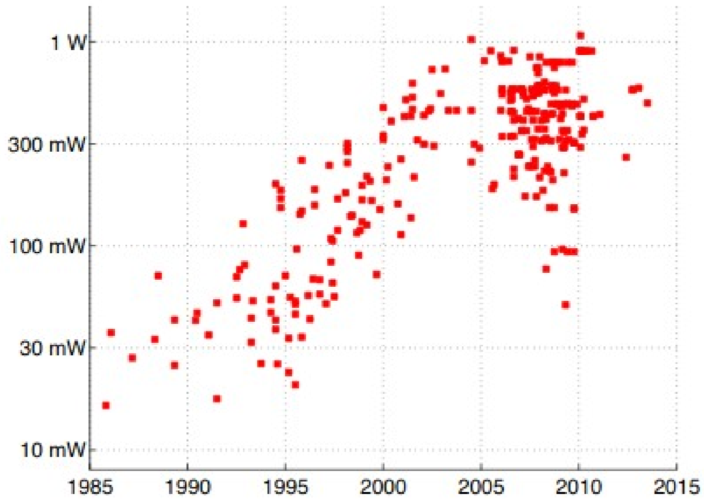


Figure 1.1.3: Power density in mW/mm² vs. year.

We hit the power wall!

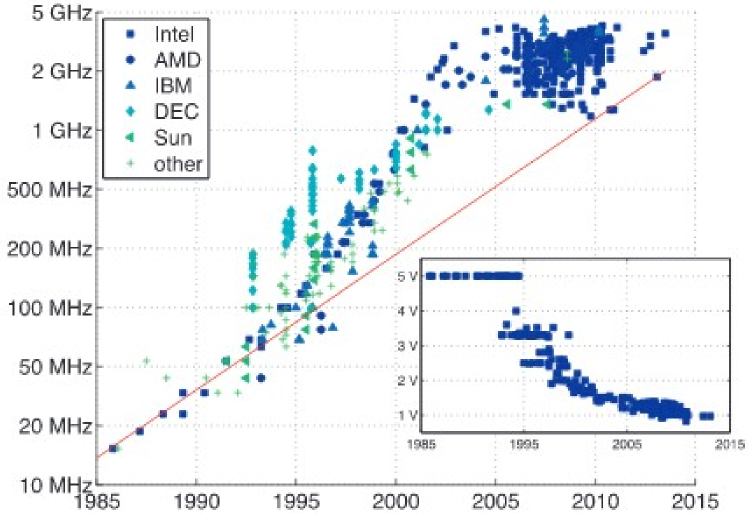


Figure 1.1.4: Clock frequency vs. year. The red line indicates frequency increase due to gate speed. The insert plot is Vdd vs. year.

Which means:
it was time for
multiple cores
per chip

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

All is well on the performance front

- Performance := Chip throughput

All is well on the performance front

- Performance := Chip throughput
- But what about the energy?

All is well on the performance front

- Performance := Chip throughput
- But what about the energy?

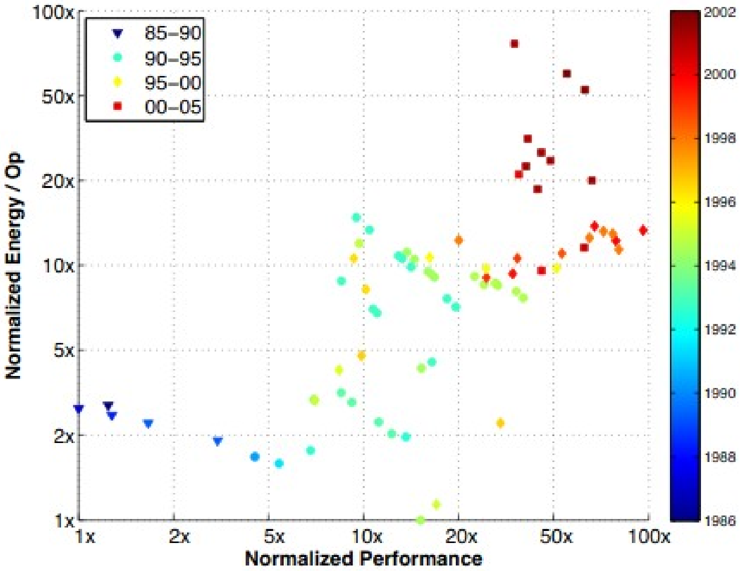


Figure 1.1.5: Instruction energy vs. peak performance (normalized).

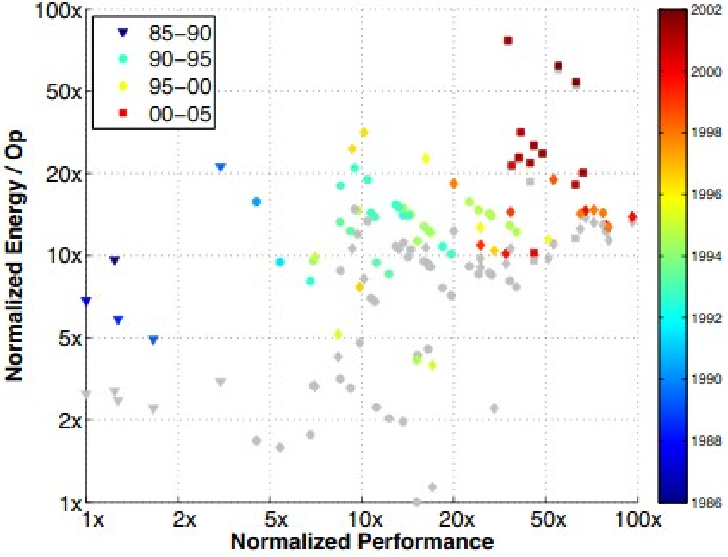


Figure 1.1.6: Instruction energy vs performance, with LLcache leakage added, with original points shown in grey for comparison.

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

It gets worse . . .

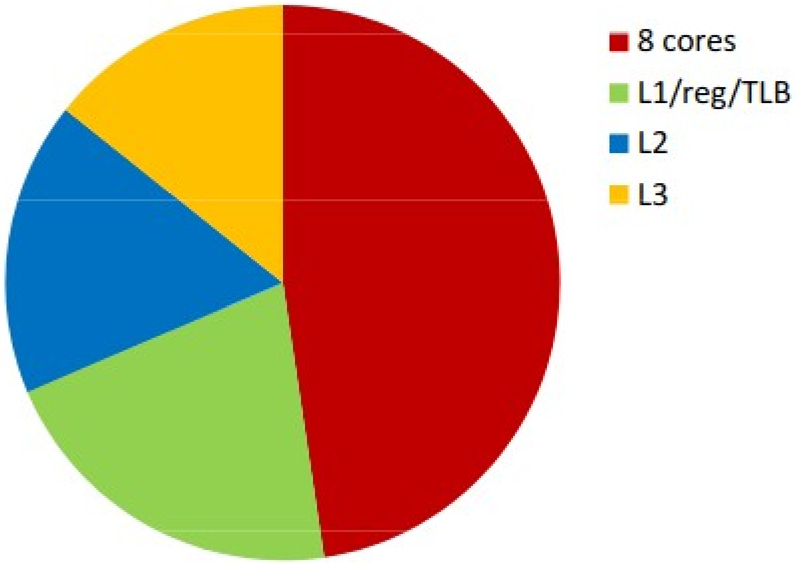
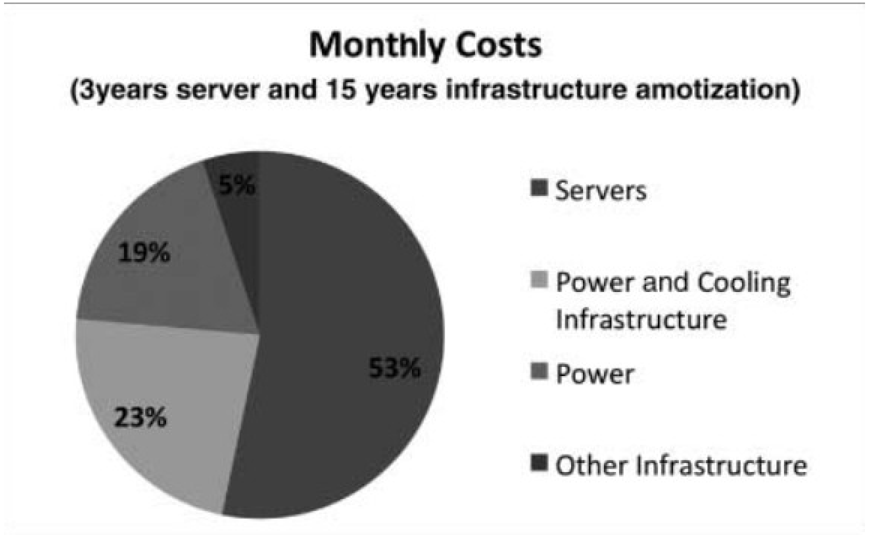


Figure 1.1.7: Power breakdown of an 8 core server chip.

Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

It gets worse . . .



Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, G., De Meer, H., Dang, M. Q., & Pentikousis, K. (2010). Energy-efficient cloud computing. The computer journal, 53(7), 1045-1051.

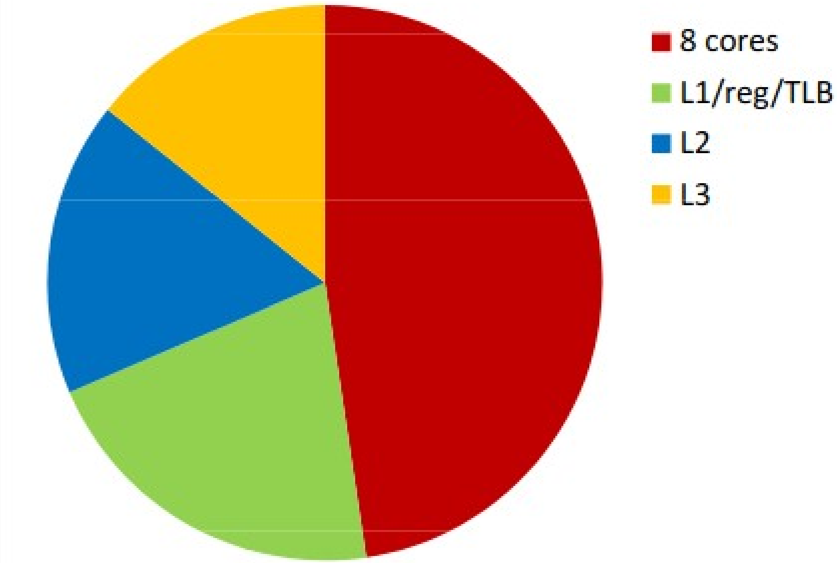
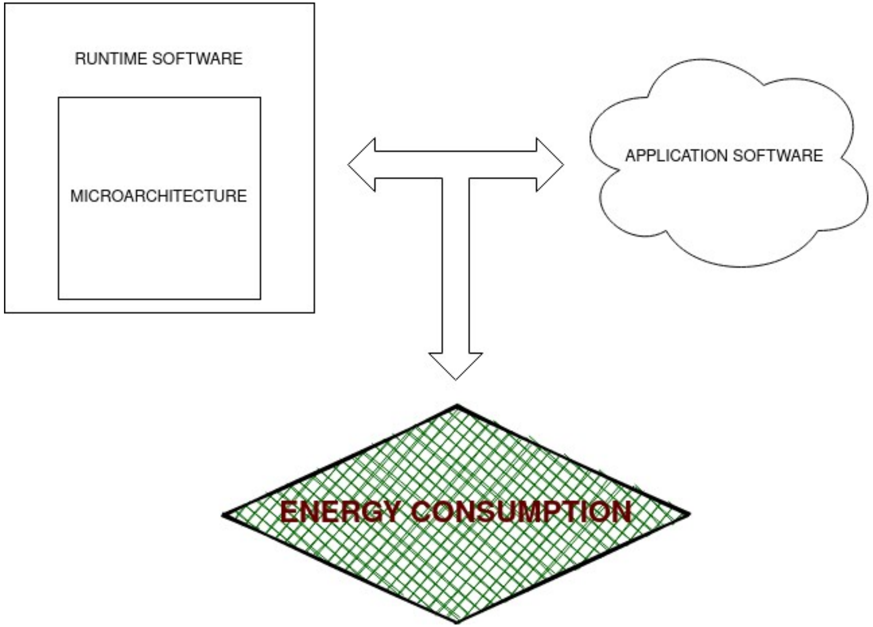


Figure 1.1.7: Power breakdown of an 8 core server chip.

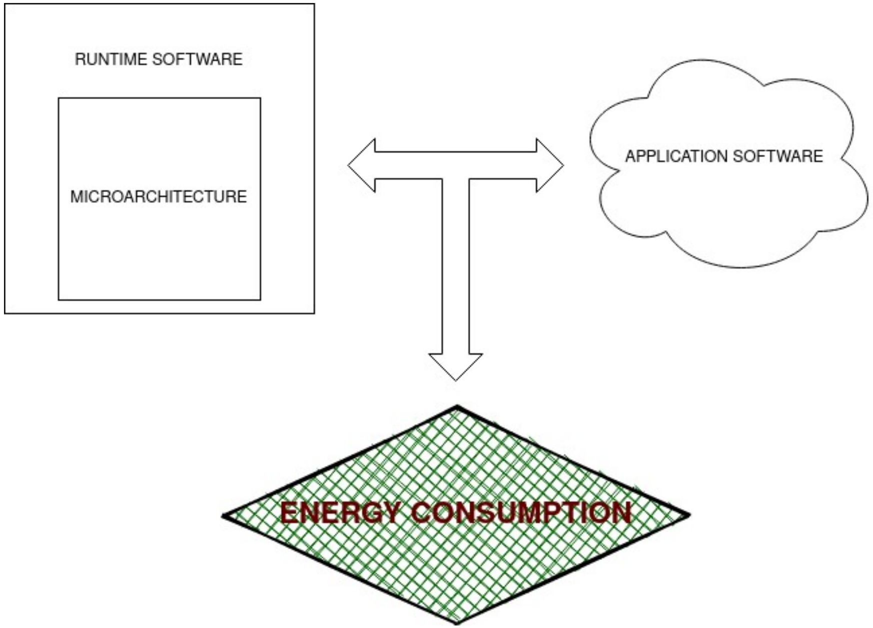
Horowitz, M. (2014, February). 1.1 computing's energy problem (and what we can do about it). In 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC) (pp. 10-14). IEEE.

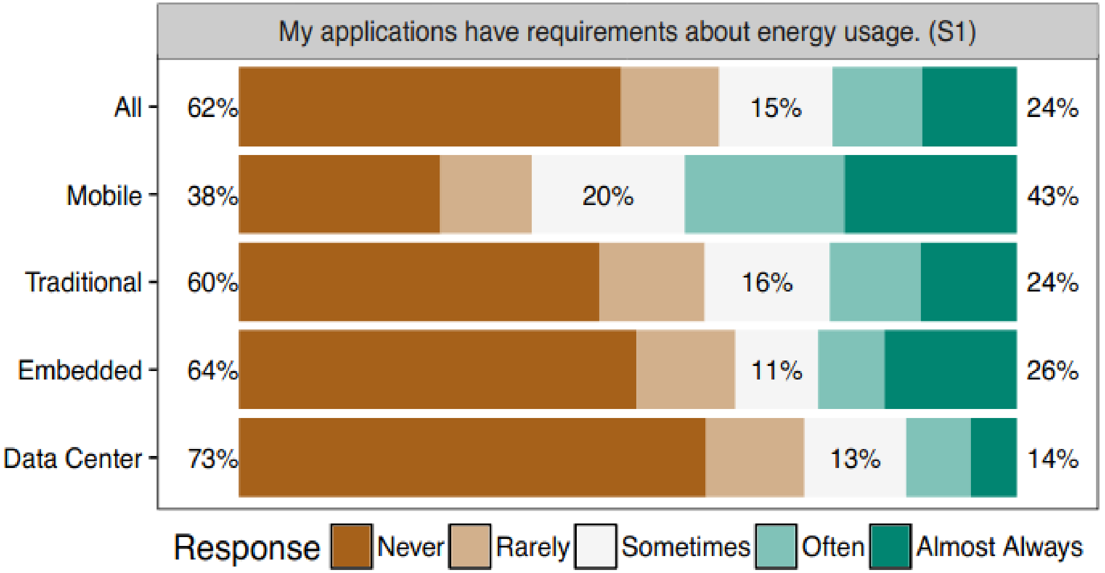
. . . and worse!



Benini, L., & Micheli, G. D. (2000). System-level power optimization: techniques and tools. ACM Transactions on Design Automation of Electronic Systems (TODAES), 5(2), 115-192.

. . . and worse!





Benini, L., & Micheli, G. D. (2000). System-level power optimization: techniques and tools. ACM Transactions on Design Automation of Electronic Systems (TODAES), 5(2), 115-192.

Manotas, I., Bird, C., Zhang, R., Shepherd, D., Jaspan, C., Sadowski, C., ... & Clause, J. (2016, May). An empirical study of practitioners' perspectives on green software engineering. In 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE) (pp. 237-248). IEEE.

So where are we?

So where are we?

- It’s complicated...

So where are we?

- It’s complicated…
  - ASIC’s (not flexible)
  - DVFS (microarch-constrained)
  - Accelerators (difficult to program)

So where are we?

- It’s complicated…
  - ASIC’s (not flexible)
  - DVFS (microarch-constrained)
  - Accelerators (difficult to program)
- ...but there are (kind of) new kids on the block!

So where are we?

RISC-V®

- It’s complicated…
  - ASIC’s (not flexible)
  - DVFS (microarch-constrained)
  - Accelerators (difficult to program)
- ...but there are (kind of) new kids on the block!

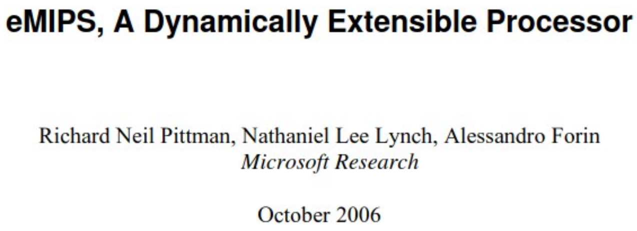
Deep Learning + RISC-V = ?

Deep Learning + RISC-V = ?

- Not much, on their own

Deep Learning + RISC-V = ?

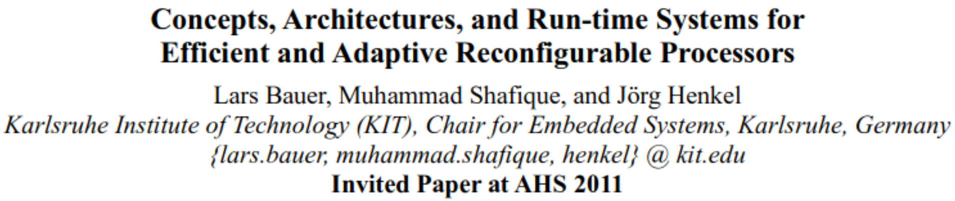
- Not much, on their own
- But what if we mixed **reconfigurable processors** in?



eMIPS, A Dynamically Extensible Processor

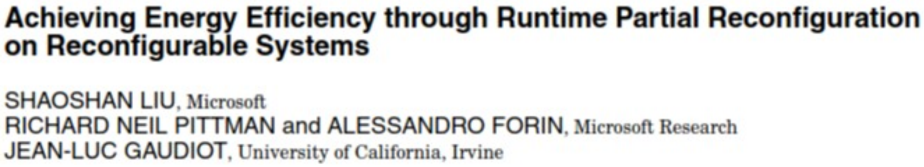
Richard Neil Pittman, Nathaniel Lee Lynch, Alessandro Forin
Microsoft Research

October 2006



Concepts, Architectures, and Run-time Systems for Efficient and Adaptive Reconfigurable Processors

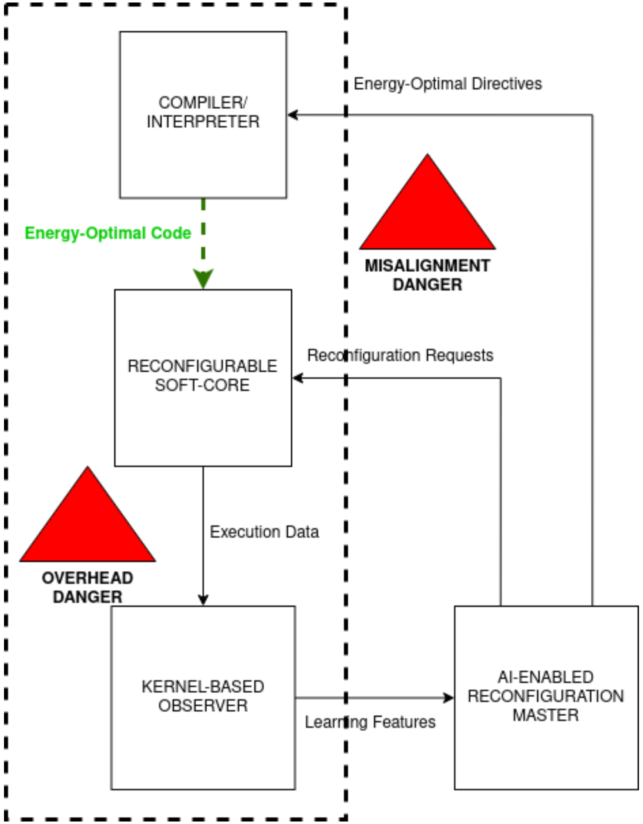
Lars Bauer, Muhammad Shafique, and Jörg Henkel
Karlsruhe Institute of Technology (KIT), Chair for Embedded Systems, Karlsruhe, Germany
{lars.bauer, muhammad.shafique, henkel} @ kit.edu
Invited Paper at AHS 2011



Achieving Energy Efficiency through Runtime Partial Reconfiguration on Reconfigurable Systems

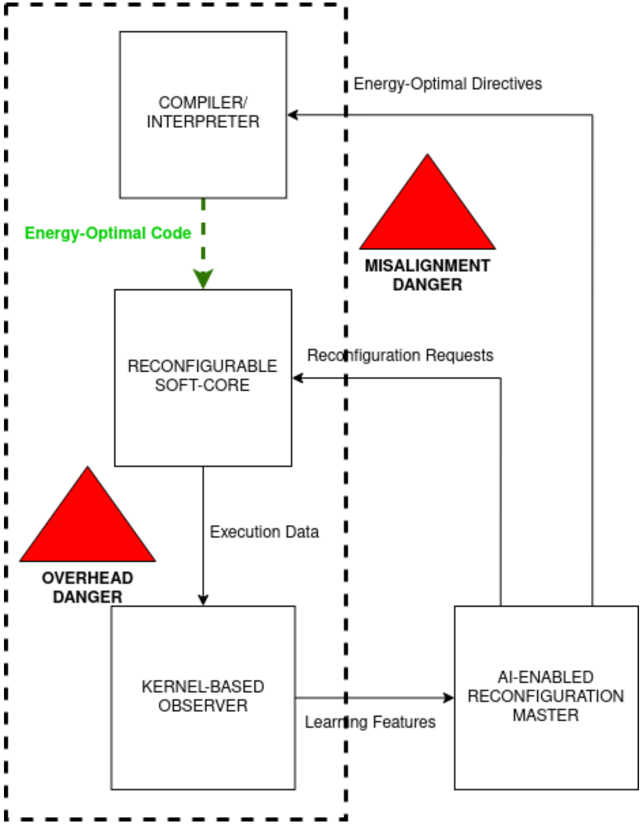
SHAOSHAN LIU, Microsoft
RICHARD NEIL PITTMAN and ALESSANDRO FORIN, Microsoft Research
JEAN-LUC GAUDIOT, University of California, Irvine

We should try this at home!

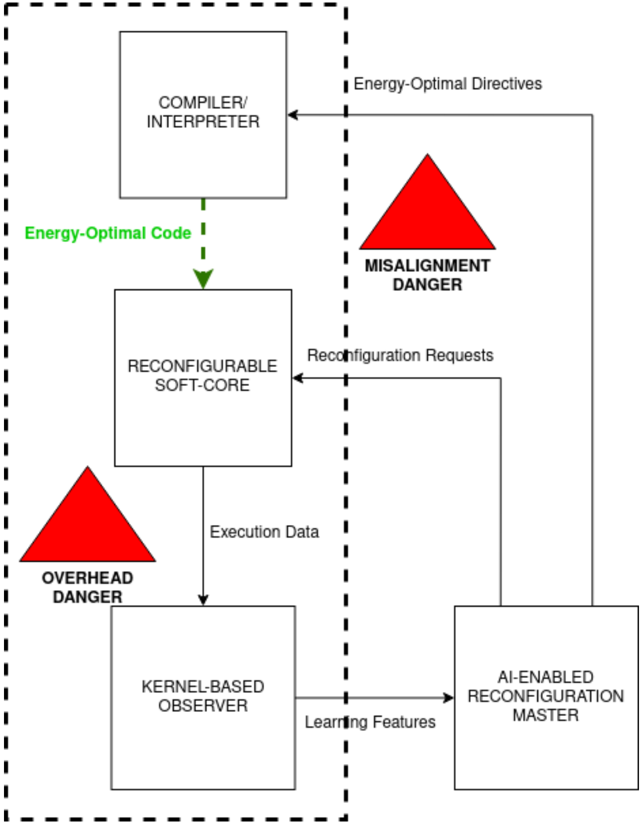


We should try this at home!

- RISC-V → reliable toolchain
- Deep Learning → complex energy models
- FPGA → dynamic adaptivity
- (Low-Power) Compiler Theory → robust background

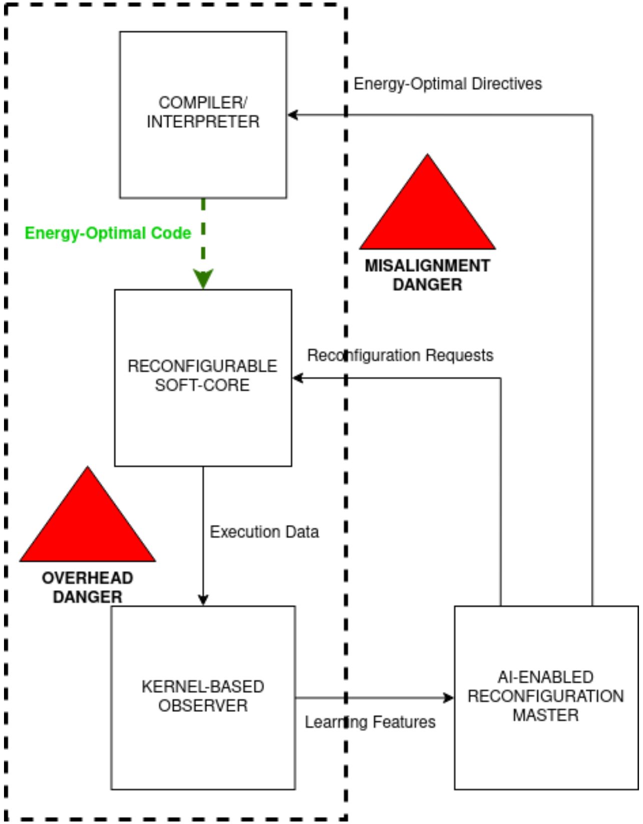


...shouldn’t we?


COMPILER/ INTERPRETER
Energy-Optimal Directives
Energy-Optimal Code
MISALIGNMENT DANGER
RECONFIGURABLE SOFT-CORE
Reconfiguration Requests
Execution Data
OVERHEAD DANGER
KERNEL-BASED OBSERVER
Learning Features
AI-ENABLED RECONFIGURATION MASTER

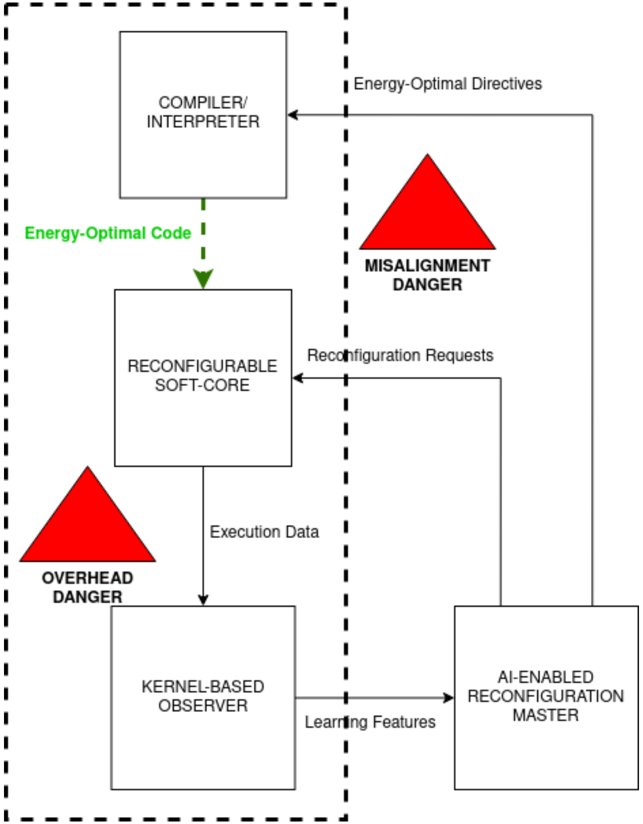
...shouldn’t we?

- *There **must** exist a reason why dynamically extensible processors haven’t conquered the industry yet*

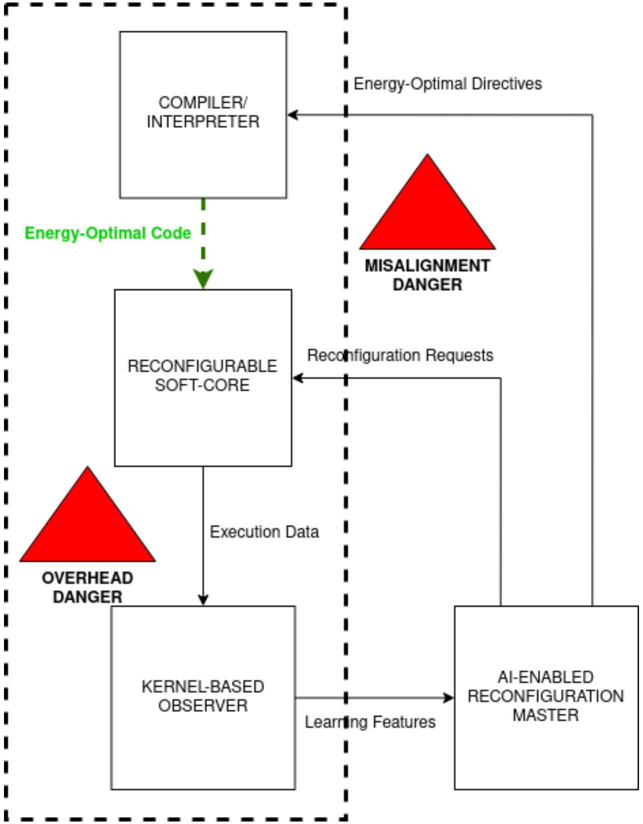


...shouldn’t we?

- *There* ***must*** *exist a reason why dynamically extensible processors haven’t conquered the industry yet*
- *Plus: the “system” imagined here has* ***a ton*** *of hidden red dangers (like the 2 shown)*

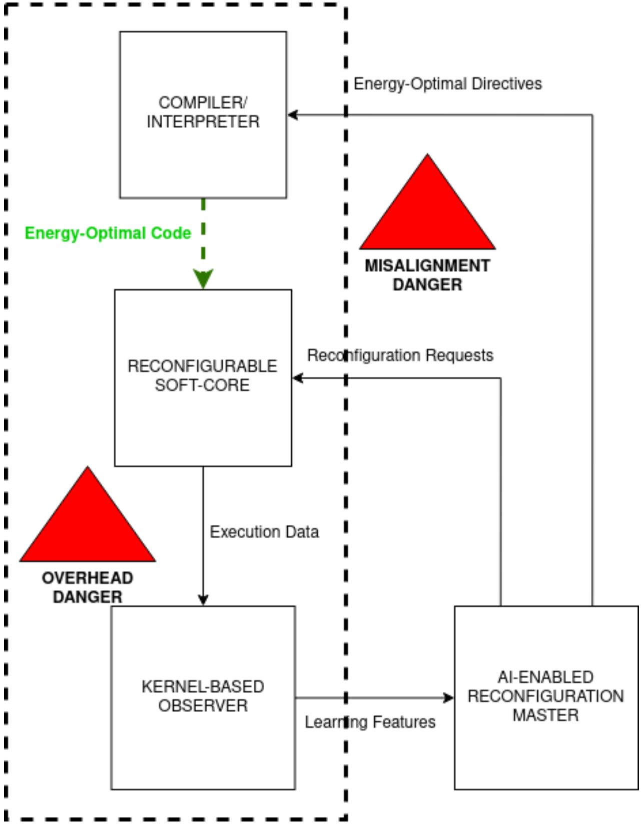


So could we go there?



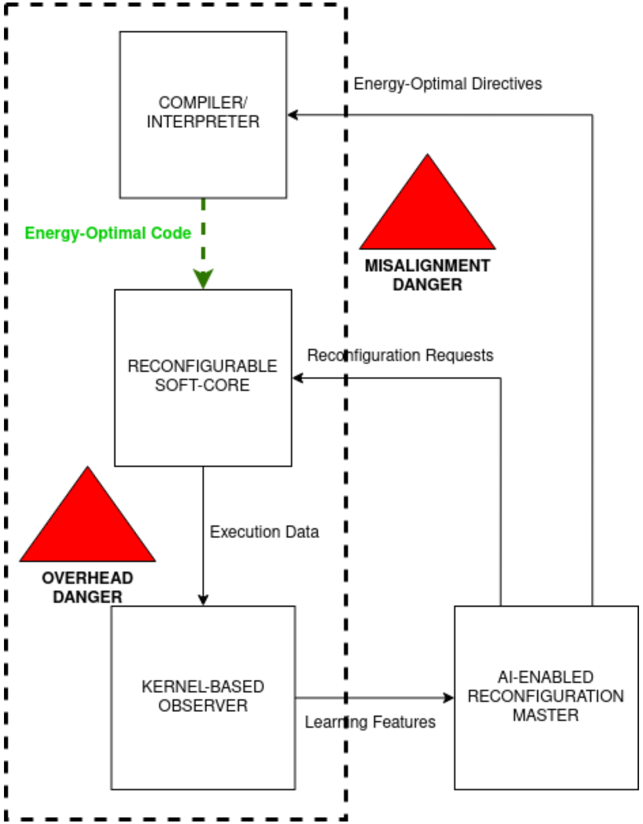
So could we go there?

- What does this look like to you?



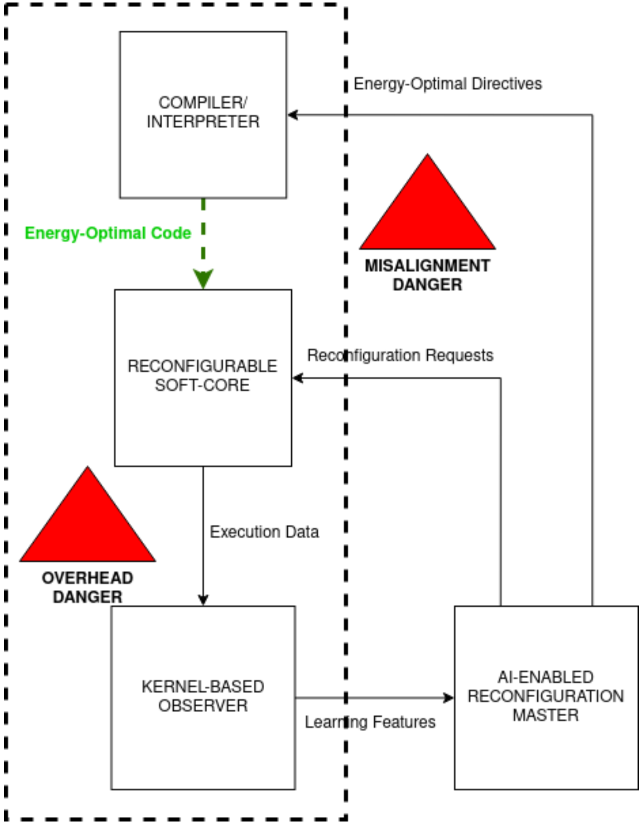
So could we go there?

- What does this look like to you?
  - A grant proposal?



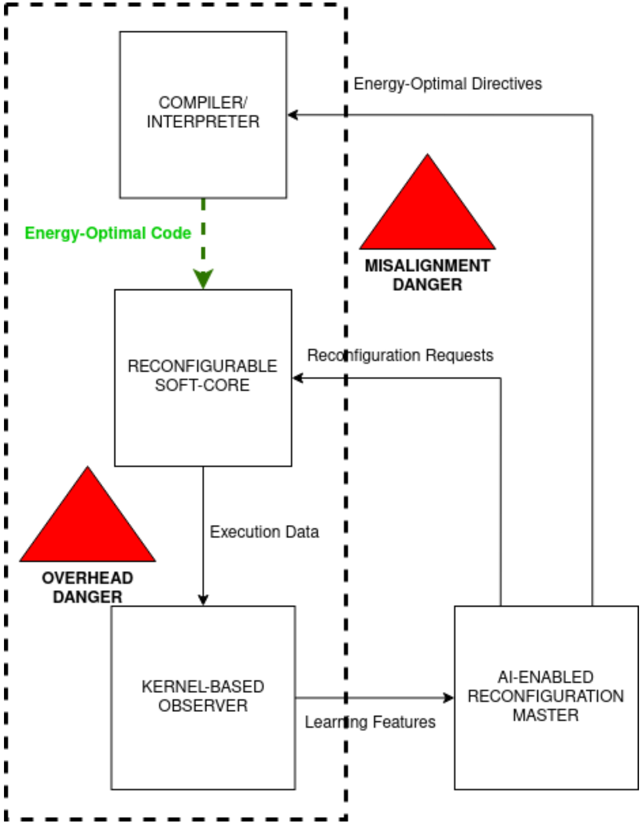
So could we go there?

- What does this look like to you?
  - A grant proposal?
  - A crazy dream of an ignorant young man?



So could we go there?

- What does this look like to you?
  - A grant proposal?
  - A crazy dream of an ignorant young man?
  - *To Mr. Soudris:* a reason to fire me?



Let’s **talk!**

THANK YOU