# Apollo: Made to measure

Andrew Pickard February 23, 2024

6 minute read time.

The ongoing evolution of technology leaves engineers with a conundrum: chips must keep getting better, but superior performance uses ever-greater power. So how do we achieve the energy-efficient CPU design the world is crying out for?

The first step is superior power management.

Back in 2020, Arm began working with Duke University on APOLLO, to develop a means of comprehensively measuring power consumption, on-chip, in real time. APOLLO's moonshot was to achieve high accuracy, even in the face of impossibly rapid power fluctuations.

Duke University's Zhiyao Xie shares how the project came about, how the team achieved fine-grain precision and how the industrial-design

cores, provided by Arm Academic Access, would prove crucial for the real-world testing of his research.

In 2020, I was a PhD student at Duke University in the United States. I interned at Arm Research, and I was assigned the task of researching power consumption.

The power consumption and performance of systems on chips (SOCs) is always a huge trade-off. Higher performance requires more power, and this limits chip development. Faster chips can only come with better power management.

Within a week or two of getting the brief, I had built a machine learning simulation tool to model the power consumption of a core. The approach showed promise, so we decided to implement it in a real circuit. It was the start of an 18-month process that taught me a great deal.

# Fine grain thinking

Power modeling is not a new topic. Many existing cores contain counters that track activities when the core is running, such as the number of caches. Power models often use those counters as inputs, taking the statistics they generate as the basis for power consumption estimates. But counters can only provide one estimation every millisecond, potentially leaving you with one estimate for millions of cycles.

> *"To ascertain per-cycle power, we would have to take a measurement every 0.3 nanoseconds. Given that one nanosecond is one billionth of a second, that's an incredibly tiny timescale."*

The thing with power consumption is that it fluctuates, often very quickly, depending on what the chip is working on. We wanted to find a way to monitor, on chip in real time, the power consumption of each

cycle. CPUs nowadays run at around three Gigahertz. To ascertain per-cycle power, we would have to take a measurement every 0.3 nanoseconds. Given that one nanosecond is one billionth of a second, that's an incredibly tiny timescale.

That June, we began experiments that would last six months or so, looking to create a fine-grain method of monitoring power. We wanted to minimize the complexity, as we did not want the module to take up too much area or resources on the CPU. So, we designed a very simple linear algorithm.

One of our key breakthroughs was how it chose which core signals to use as inputs. CPU cores are incredibly complex. They contain millions of different signals. We designed a power monitoring module that could ascertain and capture the most relevant 100. This made the model very simple, lightweight, and efficient.

With this work, we proved that you could infer hundreds of millions of cycles of power consumption within minutes. That's super-fast compared with previous power monitors.

> *"Arm Research offered a very flexible environment. It was easy to discuss questions with different people. I got a lot of support from Arm's researchers, as well as my managers."*

# Measuring success

Over the half-year, we worked on the project, I learned a huge amount. Arm Research offered a very flexible environment. It was easy to discuss questions with different people. I got a lot of support from Arm's researchers, as well as my managers. We ran a range of different experiments and covered a lot of topics. We tested on various cores, ran a hardware implementation, and created an environment with an

emulator, producing millions of cycles-worth of data. We used those results to test our algorithm.

We then went through a learning curve with paper submission, and how best to differentiate the novelty of our work from previous research. When we eventually got the paper accepted at the MICRO architecture conference in October 2021, it actually went on to be nominated for MICRO's best paper award.

# Honing the model with Arm cores

In the summer of 2022, we published an extension of the work. We improved the algorithm, and further increased the number of candidates we can select from, which made the module more flexible. But because we now had more candidates, it also made the selection more difficult. We had to propose a more complex signal selection method too, so it would perform better.

This second phase involved more engineering work, and a lot of experiments using the Arm core. Arm Academic Access gave us access to key cores and technology nodes. Having downloaded the designs and the technology library, it took several weeks to get familiar with the flow of Arm's designs. We could then produce all the layouts at the university, and we did a lot of experiments ourselves. And we presented some great results in our publications, based on the industrial-design cores that we had received from Arm.

One key difference in this work was in how we honed our selections. In the original model, we built the selection method using a pruning algorithm. We assigned weights to the millions of inputs, set penalties for each weight, and increased the penalty strength. As a result, the less important weights would shrink to zero, and you would only keep the inputs that had non-zero weights after the pruning process.

But in this model, as you penalize the weights, all of them shrink. Even those that don't shrink to zero shrink to small numbers. As a result, the pruning made the model inaccurate. The new method protected all the large weights, meaning that while some shrank to zero, others remained very large. And the model remained accurate throughout.

> *"Using Arm's cores was so different from working on toy benchmark programs… that was hugely valuable… Without the data, in fact, we couldn't even do the work."*

That pruning is now just the first step. It eliminates most of the obviously unimportant signals. We also added a further step, where the system tries different combinations to see which provides the best subset of candidates. It sweeps all the candidates and picks one to add to the selection list. It then keeps refreshing to find if there are any better.

Using Arm's cores was so different from working on toy benchmark programs. So that was hugely valuable. The data was the most important factor for us. And having that access made a massive difference to the project. Without the data, in fact, we couldn't even do the work.

The process we went through with our design was challenging. But it was rewarding too. Better measurement allows for more powerful SOCs. When we can better manage our power, the world ends up with faster chips, and who knows where that can take us?

Zhiyao Xie is an Assistant Professor at Hong Kong University of Science and Technology who completed his Ph.D at Duke University in 2022

Arm makes a wide range of IP and tools available at no charge for academic research use. To find out more, please visit our website.

Explore Research Enablement

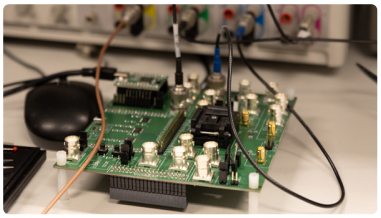
0 comments 1 member is here

## Research Articles

### Apollo: Made to measure

A researcher at Duke University developed a means of comprehensively measuring power consumption, on-chip, accurately, and in real time.

Andrew Pickard



### Processing with purpose

A research team at UCLouvain in Belgium has designed an ultra-low power Arm-based SoC that reduces the requirement for batteries.

Andrew Pickard

### An injection of ingenuity

A group of French researchers are collaborating on Arm IP to build defences against one particular type of attack, fault injection.

Andrew Pickard