

Before accelerator tape out, system developers ensure hardware-accelerated full-stack is performant  
Their typical **workflow** is:

1. Setup end-to-end simulator

2. Offload tasks to the accelerator
3. Simulate the full-stack

⌚ Wait for hours

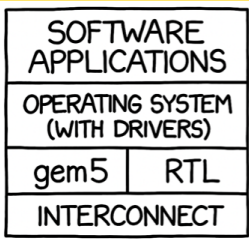
4. Get profiling result, e.g., timer values

5. ⚠ Doesn't look good

6. Tune the stack 🔧

7. For the next few weeks or months

• Repeat step 3 - 6



Toolbox 🔧

- Increase async execution
- Reduce contention
- Pipelining
- Batching
- Change accelerator config
- ...

Developers are forced into **non-interactive** workflow, which hurts **productivity**

**Root cause** of slowness is gem5+RTL provides too much **visibility**

