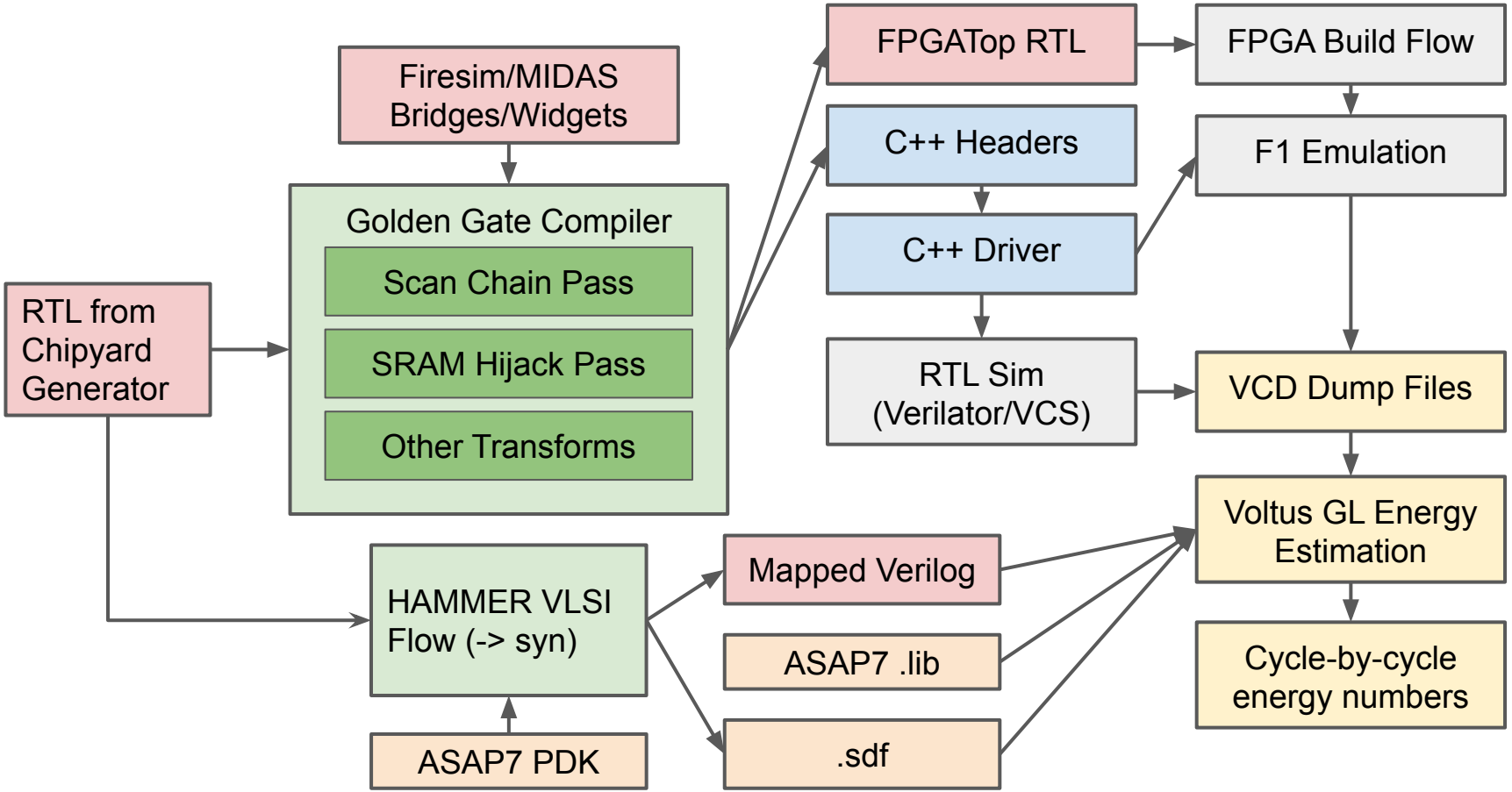
Energy Estimation Flow



Checkpoint 2 (4/27) Goals and Progress

- RTL simulation of a simple circuit (Risc/GCD) after default Firesim transformation
- ASAP7 Genus synthesis of the circuit and SRAM macros
- Formal mapping of RTL VCD to gate-level VCD
- GCD/Decoder Power Estimation using Voltus vectored analysis
  - Working through some Voltus issue where it can't read the Innovus DB properly, the top cell name isn't being exported by Innovus, and we're trying to find out why
- Rocket-chip w/ Gemmini running matmul in local RTL sim -> energy estimation using Voltus vectored analysis
  - Voltus/HAMMER running in Chipyard environment
  - Power estimation chipyard branch exists, we're trying to use the Makefile appropriately
- Shadow scan-chain instrumentation pass + SRAM hijack pass
  - MIDAS-level VCD dumping RTL simulation

Changes to Timeline

- Newly refined goals are now split between:
  - 1) energy estimation from RTL simulation and Voltus GL sim on small GEMMs and other operations like pooling and dwconv. Using these to estimate the power of a full DNN execution by its constituent parts (e.g. mobilenet v2)
    - Get FU-level energy estimates (replicate Horowitz's plot on energy/MAC, energy/SRAM R/W, energy/DRAM access, etc.) Observe patterns for density too. Create a rough architectural-based energy estimate and cross-check with Voltus.
  - 2) getting MIDAS-level RTL sim of the scan chain pass working for small modules (maybe not a full rocket-chip w/ Gemmini), and being able to dump VCDs or **SAIFs** from simulation
- 1) We can do this by running many small **GEMM/dwconv/pooling/im2col ubenchmarks** in parallel and sending each one to Voltus in parallel and combining the results later. This should give us a good estimate of Gemmini's energy efficiency for a full DNN if the FPGA sim doesn't work. Test **different input matrix sparsities** to simulate the impact on Gemmini's efficiency/clock gating.
- 2) If dumping a full VCD proves too inefficient, we can design some FPGA RTL that converts raw scan chain data to SAIFs and dump that instead. Instrumentation of Gemmini for MAC utilization.
- Final report expectation is data for DNNs from part (1) and a set of FIRRTL passes for (2) that can be finished later for full Strober-like functionality