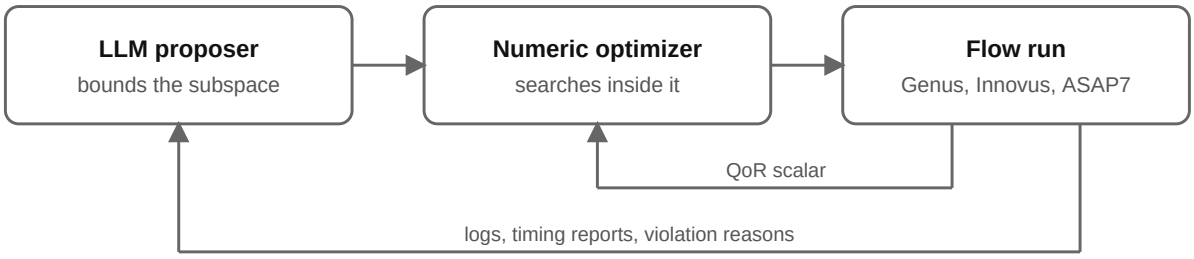
# A Hybrid Flow Tuner: LLM Proposer over a Numeric Optimizer

EECS 6693 design project · Wassim Kassem (wkk2104) and Xingcan Chen (xc2807) · October 11, 2026

Flow tuners optimize against a single scalar, but the tools emit far more: logs, timing reports and stated reasons for each violation, none of which a numeric optimizer can read. We propose a tuner in which a language model reads that output and bounds the region of the parameter space worth searching, and a conventional optimizer searches inside that region. The two have no overlapping role.




*The asymmetry is the contribution: the numeric optimizer sees only the scalar, the LLM sees why a run failed.*

### What gets built

A tuning harness around the Genus / Innovus / ASAP7 reference flow with a pluggable proposer interface, so every arm runs through the same evaluation path and budget accounting. The LLM proposer consumes the run's reports and returns a bounded subspace plus a rationale; the numeric optimizer (Bayesian optimization, or a multi-armed bandit) searches within it.

### How it is evaluated

All arms are compared at matched run count and wall-clock, since a proposer that spends more compute proves nothing. Four arms: random search; the numeric optimizer alone; a flat portfolio where both propose and a bandit splits the budget; and the hierarchy above. The flat portfolio is the ablation isolating whether the hierarchy matters or merely the presence of an LLM.

Reported separately: best QoR per CPU-hour against each baseline; the LLM's win rate when it competes for budget in the flat arm; and how often its stated rationale matches the violation the next run actually resolved. If the LLM never wins budget, that is a result and we report it.

### Why this is not two projects stapled together

Remove the LLM and this is a conventional tuner, with dense prior art (SynTunSys, FlowTune, DRiLLS, OpenROAD AutoTuner). Remove the numeric optimizer and the LLM is doing unguided search in a continuous space, which it is bad at. The claim is that semantic bounding plus numeric search beats either alone at equal cost.

### Risks

**Compute.** Four arms across several designs is the binding constraint; design count follows the parallel run budget. **Variance.** Model nondeterminism means repeated seeds and reported spread per arm, not single numbers. **Scope creep.** The LLM stays a proposer inside a fixed action space, not a general agent editing the flow.

### To confirm before the proposal

How many parallel Genus and Innovus runs the class servers sustain; whether tuners from previous 6693 projects exist in runnable form and may be reused as baselines; and which designs are available as a shared benchmark set.

---

*Combines idea A (automated PPA flow tuner, X. Chen) and idea H (LLM-driven flow repair agent, W. Kassem) from the team's candidate list of October 5.*