

Search space & hardware feedback

EvoTuner extends Triton’s manual configuration list with model-guided evolutionary search. It benchmarks promising candidates and keeps the fastest measured configuration.

1 Configuration as chromosome

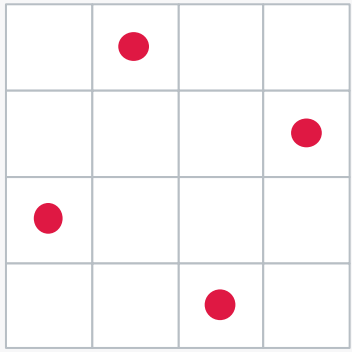
One triton.Config becomes one chromosome. Each gene stores its name, value, bounds, and metadata flags. Seed-wise minima and maxima define the search bounds.

| Gene group         | Examples              | Allowed values   |
|--------------------|-----------------------|------------------|
| Kernel parameters  | BLOCK_M / N / K       | Powers of two    |
| Execution metadata | num_warps             | Powers of two    |
| Execution metadata | stages, CTAs, threads | Bounded integers |

Clamp repairs values after variation. Sorted (name, value) signatures identify duplicates. Conversion restores kernel kwargs and backend execution options.

2 Latin Hypercube Sampling (LHS)

If the discrete space fits the generation budget, enumerate it. Otherwise generate N + W samples: N for the population and W for the warmup pool.



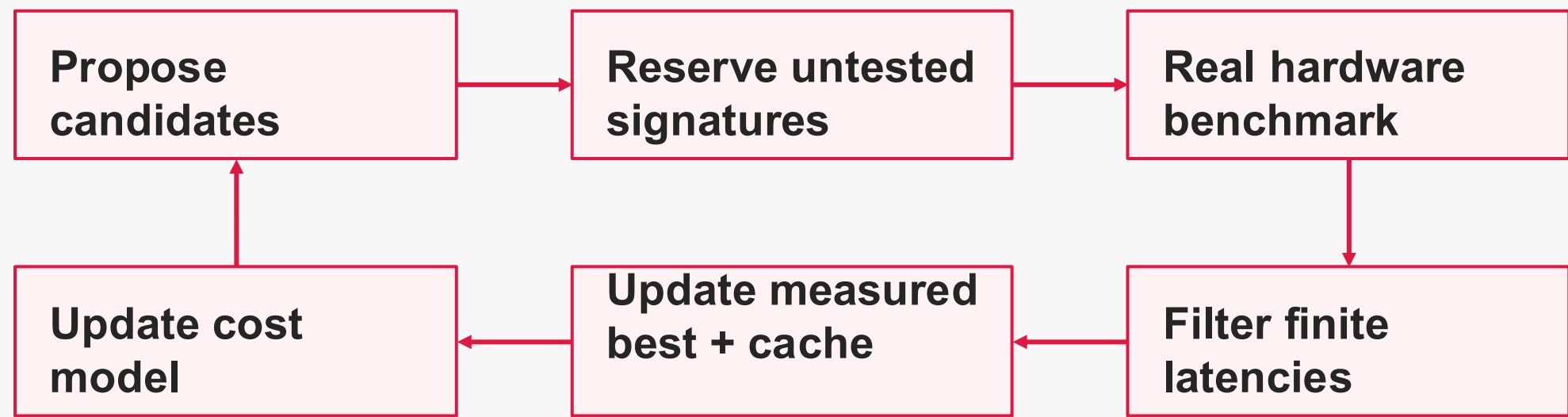
4-sample example

LHS generates diverse initial configurations by spreading samples across each parameter’s range without enumerating every combination.

For each dimension, divide [0, 1) into equal intervals, sample once per interval, and shuffle the samples independently. Convert each sample u into a legal parameter value using the index floor(u × m), where m is the number of legal values. For example, with tile sizes [16, 32, 64, 128], u = 0.6 selects 64.

3 Learned cost model and real feedback

Features: configuration values plus each candidate’s resolved grid\_0, grid\_1, grid\_2. The XGBoost model learns from finite measured latency and returns search scores.



Benchmark median latency with quantiles (0.5, 0.2, 0.8). Failed configurations receive infinite latency and do not train the model. Re-score the population after each update.

Experiment

Evaluation: Ryzen 9 9950X CPU and RTX 3060 Ti GPU. AUTO and EVO share the kernel body and initial list. Population 40, generations/round 30, top-k 3.

Budget 150, warmup 45 (LayerNorm: 90 / 27).  
Geomean speedup = AUTO latency / EVO latency.

| Kernel          | CPU    | GPU    |
|-----------------|--------|--------|
| Blocked matmul  | 1.195× | 1.024× |
| Batched matmul  | 1.691× | 0.971× |
| Dequantization  | 0.977× | 0.990× |
| Fast LayerNorm  | 1.036× | 0.981× |
| Fused attention | 1.688× | 0.984× |

CPU batched matmul wins 19/19 shapes. Fused attention wins 15/16 and reaches 2.789× maximum speedup. GPU gains are smaller and inconsistent.

Evolution & adaptive strategies

4 Selection: choose the parents

Roulette-wheel sampling draws pop\_size parents with replacement from cost-model probabilities. Promising candidates are more likely to reproduce, while others retain a chance.

5 Uniform crossover: recombine genes

Pair copied parents. The pair’s crossover rate and per-gene swap probability are the averages of both parents’ strategy parameters. Swap aligned genes independently, then clamp.

Illustrative recombination (values are not benchmark results)

| Candidate         | M  | N  | K  | warps |
|-------------------|----|----|----|-------|
| Parent A          | 16 | 32 | 64 | 4     |
| Parent B          | 64 | 64 | 32 | 8     |
| Child A (A,B,B,A) | 16 | 64 | 32 | 4     |

No crossover: keep parent copies. An unpaired parent also passes through as a copy. Mutation still follows.

6 Mutation: choose a search step

| Strategy     | Gene mutation | Pair crossover | Gene swap |
|--------------|---------------|----------------|-----------|
| Exploration  | 0.35          | 0.75           | 0.55      |
| Exploitation | 0.12          | 0.85           | 0.35      |
| Random       | 0.70          | 0.20           | 0.80      |

**Exploration:** Explore new configurations with larger mutations. Integer metadata uses Gaussian perturbations with  $\sigma = \max(1, 0.25 \times \text{range})$ . Power-of-two genes use Gaussian perturbations in  $\log_2$  space with  $\sigma = 1.5$ .

**Exploitation:** Refine current configurations with smaller mutations. Integer metadata uses  $\sigma = \max(1, 0.08 \times \text{range})$ , while power-of-two genes use  $\sigma = 0.5$  in  $\log_2$  space.

**Random:** Replace each selected gene with a value sampled uniformly from its legal integers or powers of two.

Here, range = upper bound – lower bound.

7 Family competition and strategy change

Compare each parent and child using model scores. Keep the better one; keep the child on a near tie ( $|\Delta \text{score}| \leq 10^{-6}$ ). Then sample the survivor’s next strategy.

Fixed transition probabilities depend on the parent’s strategy and the predicted win / loss / tie. They are not learned from benchmark wins.

Next-strategy probabilities

| Parent strategy | Outcome | Explore | Exploit | Random |
|-----------------|---------|---------|---------|--------|
| Exploration     | Win     | 25%     | 65%     | 10%    |
| Exploration     | Loss    | 70%     | 15%     | 15%    |
| Exploration     | Tie     | 50%     | 25%     | 25%    |
| Exploitation    | Win     | 10%     | 80%     | 10%    |
| Exploitation    | Loss    | 60%     | 25%     | 15%    |
| Exploitation    | Tie     | 35%     | 50%     | 15%    |
| Random          | Win     | 25%     | 45%     | 30%    |
| Random          | Loss    | 45%     | 15%     | 40%    |
| Random          | Tie     | 30%     | 20%     | 50%    |

8 Candidate pools, rounds, and execution



After each generation, retain untested candidates in Top-k (up to k) and Frontier (up to 5k) pools. The arrows show selection priority: choose from Top-k first, then fill remaining slots from Frontier and the ranked population.

Deduplicate and reserve signatures before benchmarking up to k candidates. New tuning keys reset the search and model, while cache hits reuse the fastest measured configuration.

Acknowledgement

This work is supported in part by PeakHills Group, MediaTek.