



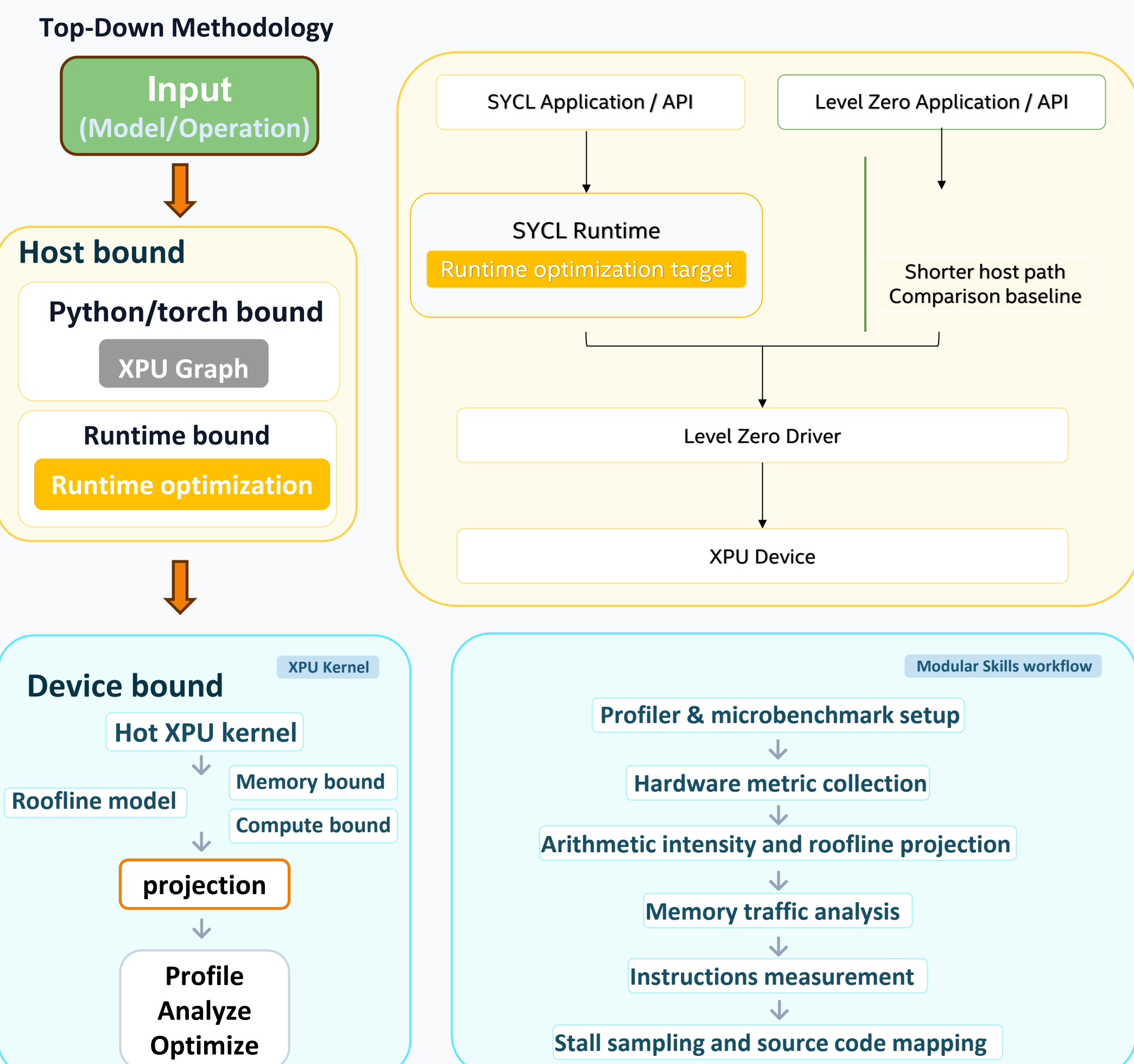
BACKGROUND

- 🕒 We introduce a top-down performance analysis workflow to identify inefficiencies from model-level down to low-level runtime, kernel, and hardware behavior.

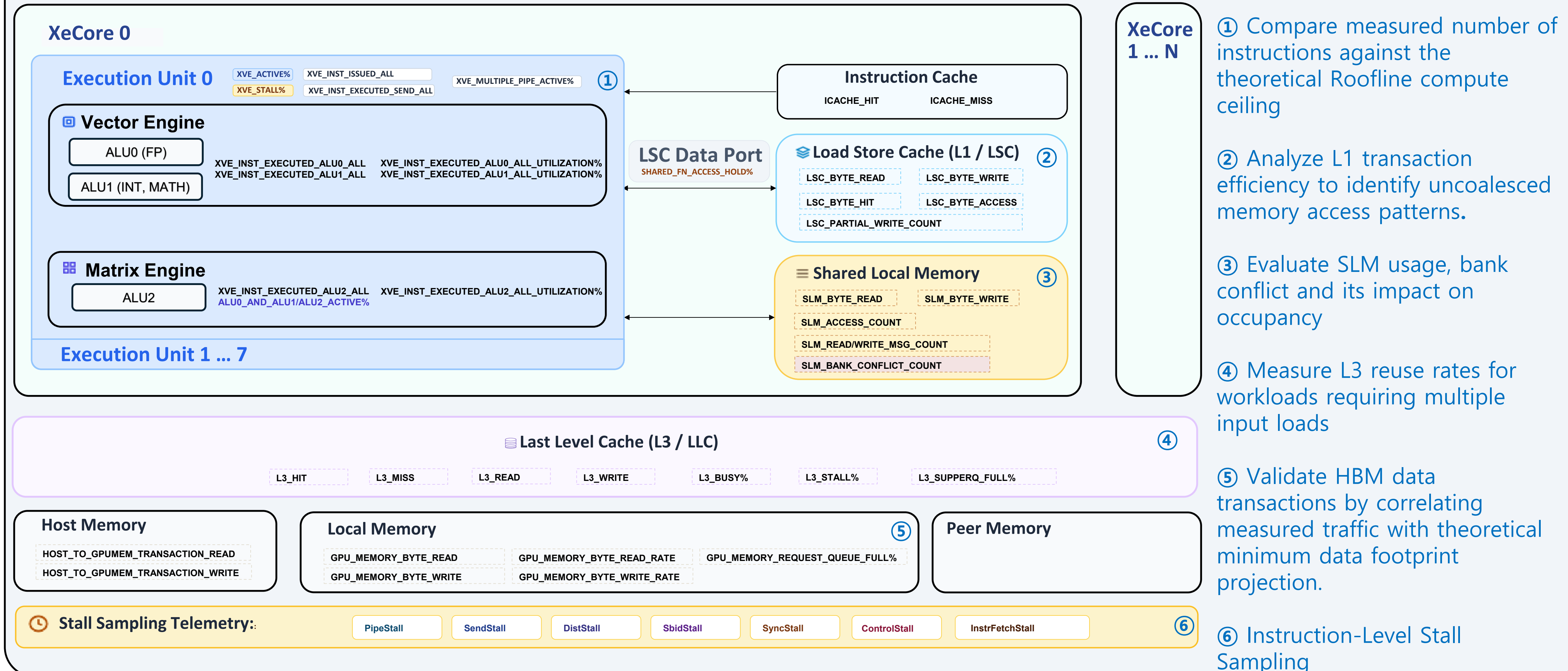
The Cross-Layer Software Stack



END-TO-END DECISION FLOWCHART



DEVICE OPTIMIZATION WORKFLOW



- XeCore 1 ... N**
- Compare measured number of instructions against the theoretical Roofline compute ceiling
 - Analyze L1 transaction efficiency to identify uncoalesced memory access patterns.
 - Evaluate SLM usage, bank conflict and its impact on occupancy
 - Measure L3 reuse rates for workloads requiring multiple input loads
 - Validate HBM data transactions by correlating measured traffic with theoretical minimum data footprint projection.
 - Instruction-Level Stall Sampling

HOST OVERHEAD OPTIMIZATION

