Let’s run a million benchmarks

Yao Yue, IOP Systems

When to run O(1) benchmarks

“Just upgraded my SSD, let me make sure this thing works.”

When to run O(10) benchmarks

“What should my next desktop be to play my favorite game?”

When to run O(100) benchmarks

“How well does my service work for all the customer workloads I have?”

When to run O(1k) benchmarks

“How do I best configure my mission-critical service A for a given workload?”

One million benchmarks is not that far away...

“How do I customize application configuration for each of all the customer workloads for best bang for the buck?”

“How can we survey as many popular projects as possible to design the next-gen hardware?”

“Let’s run regression / optimization benchmark suite nightly.”

Coverage, combination, time

The Logistics of 1 Million Benchmarks

Some napkin math...

How long does it take?

- Usually 1 to 60 minutes per test => up to 1M machine hours

How much data does it produce?

- Assuming 10KB - 100MB per test => 10GB to 100TB

RADBOI! (Rub A DataBase On It)

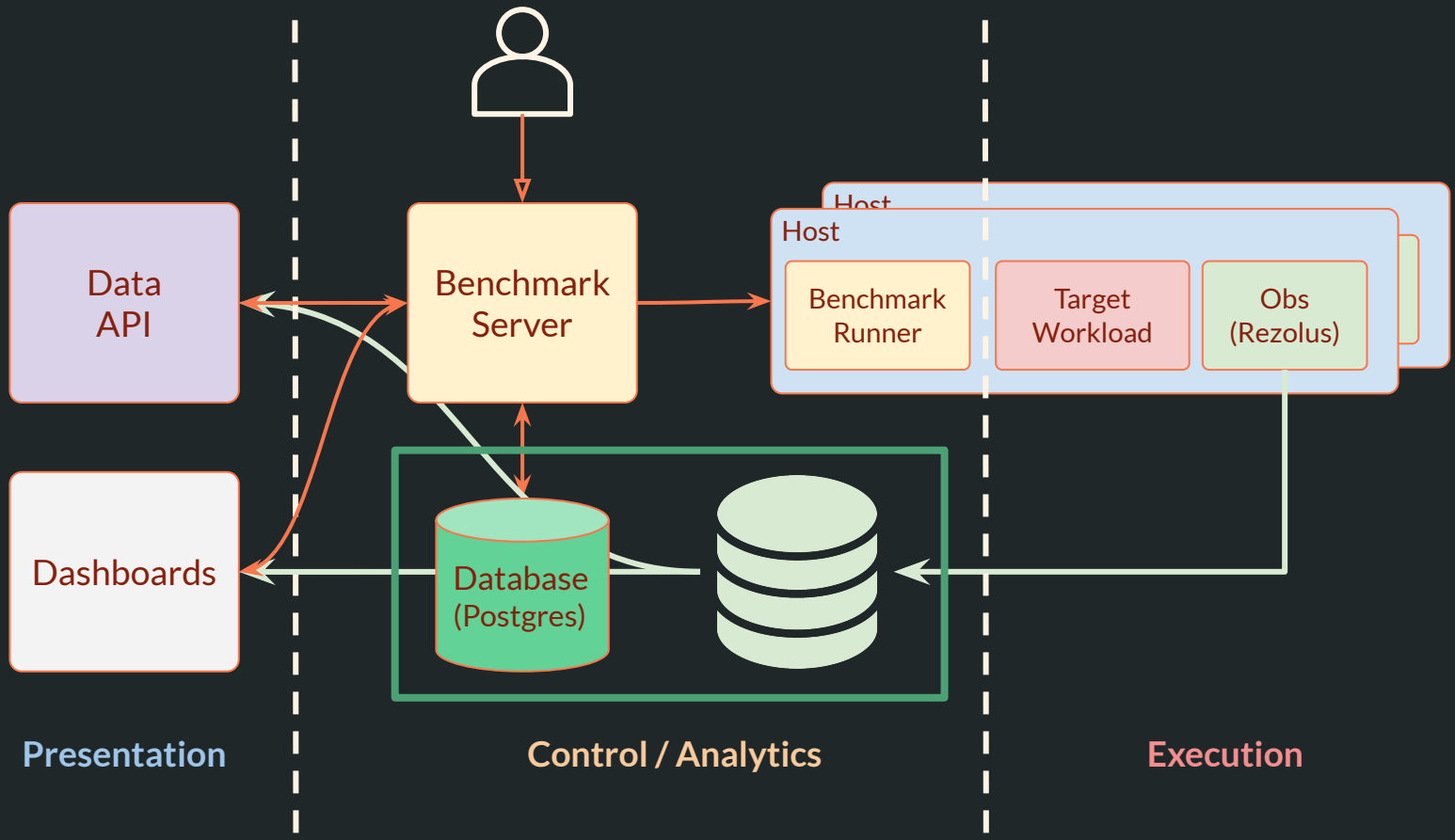
To manage benchmark execution

Automation!

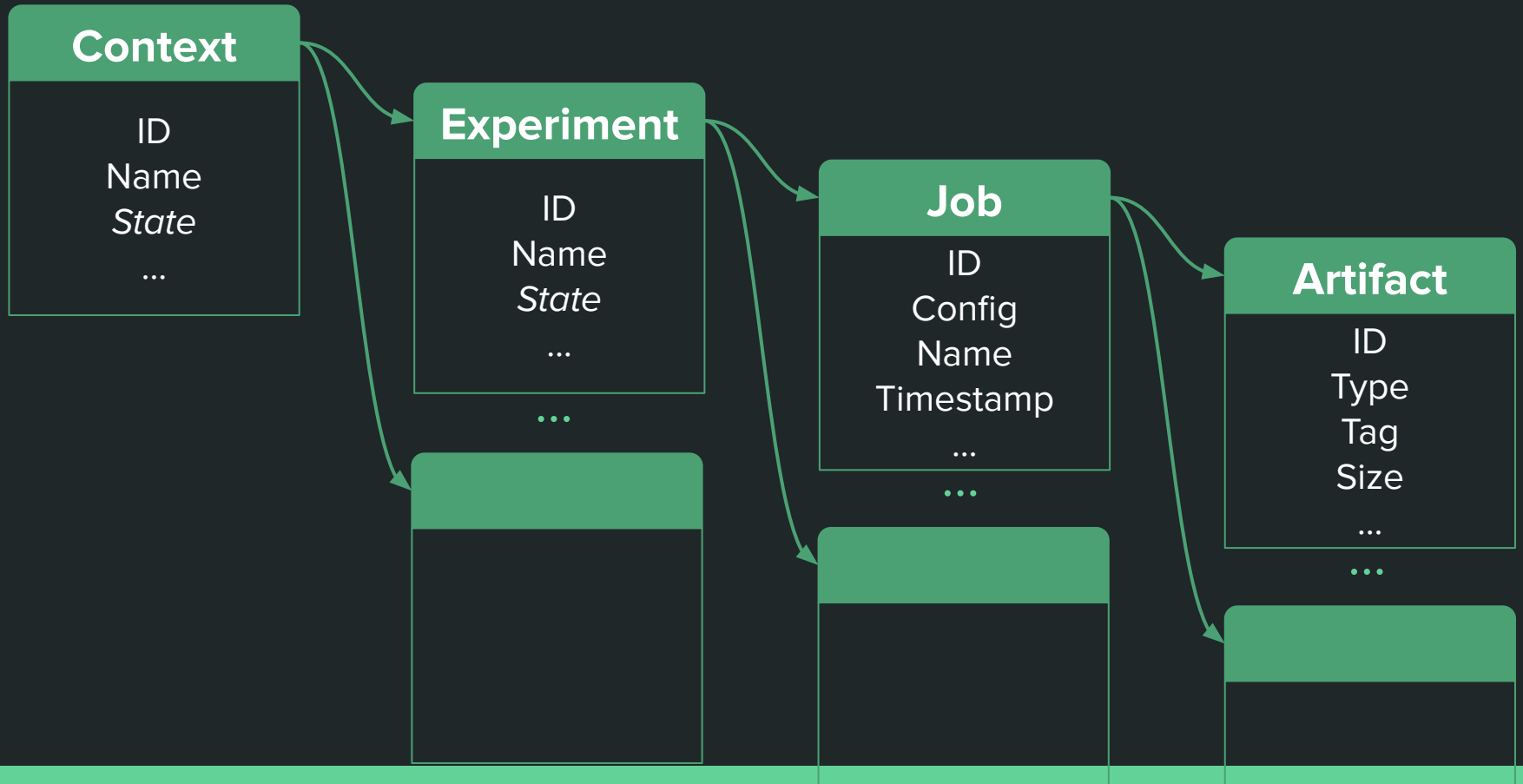
To organize and distill benchmark results

Analytics!

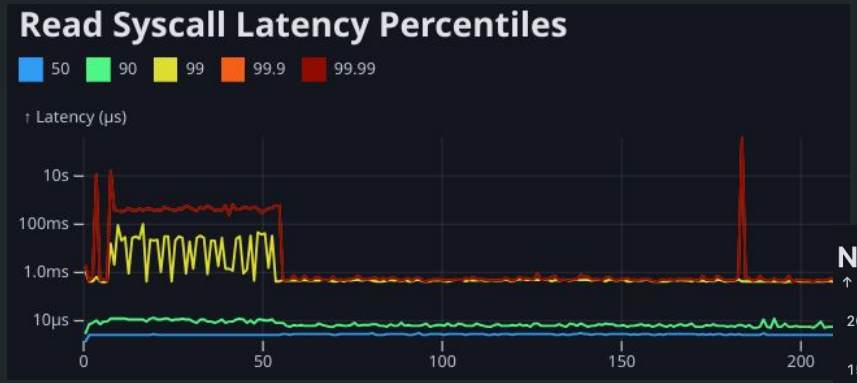
Fitting it in

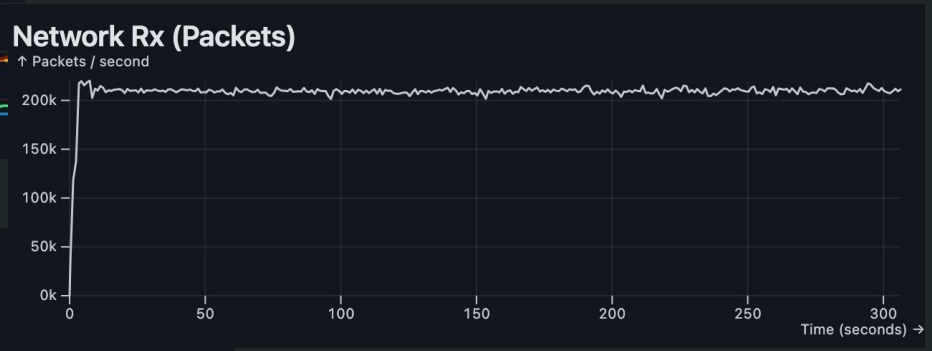


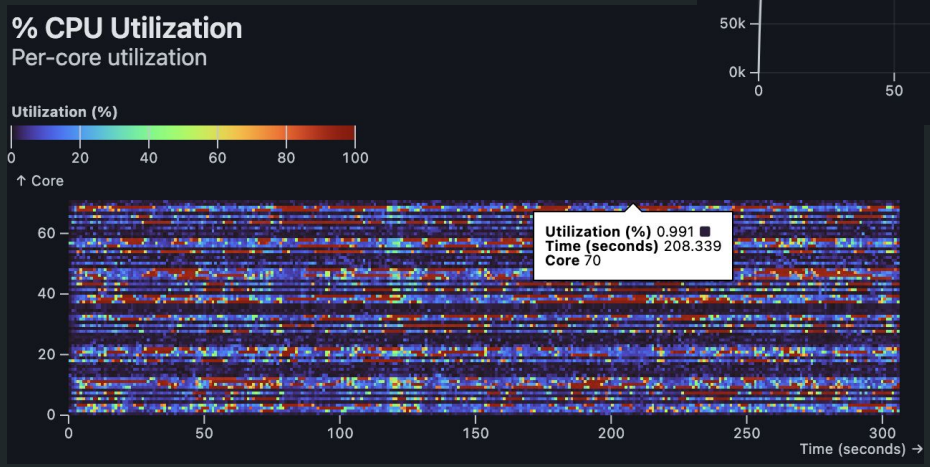
What's in the database



Database enables auto-collection & processing…







In a world where running 1 million benchmarks is easy, it becomes **a tool for insight discovery**.

Database makes it easy to slice and dice...

Parameter Name	Baseline Value	Best Candidate Value
ec2	m6i.xlarge	m7i.xlarge
jdk	jdk8	**jdk11**
🔒 message_size	512	512
🔒 tls	plain	plain
🔒 compression_ratio	1	1
compression	lz4	**zstd**
linger_ms	0	**5**
batch_size	16384	**524,288**
key_size	8	8

Measurement	Baseline Value	Best Candidate Value
throughput	39.9k	249k
cpu	79.3%	68.3%

Database enables sweeping summaries.

Dollar Cost per Million Tokens

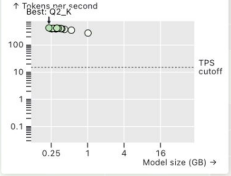
Cost per 1M tokens ($)

0.2 1 2 10

	135M SmolLM	360M SmolLM	1B Llama 3.2	1.7B SmolLM	2B Gemma 2	3B Llama 3.2	7B Mistral v0.3	9B Gemma 2	12B Mistral Nemo	13B Llama 2	27B Gemma 2
g6e.xlarge	1.0	1.2	1.0	1.3	2.1	2.0	3.0	4.6	4.7	4.8	9.6
g5g.xlarge	0.4	0.5	0.7	0.8	1.5	1.5	2.8	3.9	4.0	4.3	×
c7i.xlarge	0.4	1.0	2.2	2.8	×	×	×	×	×	×	×
c6i.xlarge	0.4	0.8	1.8	2.5	×	×	×	×	×	×	×
c6a.xlarge	0.2	0.5	1.2	1.6	2.8	2.8	×	×	×	×	×
c7a.xlarge	0.2	0.4	1.0	1.2	2.3	2.3	×	×	×	×	×
c6g.xlarge	0.2	0.3	0.9	1.1	2.2	2.1	×	×	×	×	×
c8g.xlarge	0.1	0.2	0.7	0.8	1.7	1.6	2.9	×	×	×	×
c7g.xlarge	0.1	0.2	0.7	0.8	1.6	1.6	×	×	×	×	×
g5.xlarge	0.8	0.9	0.7	1.0	1.6	1.5	2.4	3.6	3.7	3.9	7.7
g5g.2xlarge	0.6	0.8	1.0	1.3	2.3			7.0	×	×	×
c6i.2xlarge	0.4	0.9	2.0	2.6	4.8			×	×	×	×
c6a.2xlarge	0.3	0.6	1.5	1.8	3.4			×	×	×	×
c7a.2xlarge	0.3	0.8	1.9	2.2	4.3			×	×	×	×
c7i.2xlarge	0.5	1.0	2.3	2.9	5.0			×	×	×	×
c7g.2xlarge	0.2	0.4	0.8	1.0	2.0			5.3	×	×	×
c8g.2xlarge	0.2	0.3	0.8	1.0	1.9			5.2	×	×	×
c6g.2xlarge	0.3	0.4	1.1	1.2	2.7			×	×	×	×
c6i.4xlarge	0.6	1.2	2.5	3.2	6.1			×	×	×	×
c7i.4xlarge	0.5	1.3	2.6	3.3	5.9			×	×	×	×
c7a.4xlarge	0.6	1.1	2.2	2.7	5.3	4.9	9.1	13.9	×	×	×
c6a.4xlarge	0.5	1.1	2.6	3.2	6.3	6.2	×	×	×	×	×
c6g.4xlarge	0.5	0.7	1.5	1.7	3.8	3.1	5.9	9.6	×	×	×
c8g.4xlarge	0.4	0.6	1.1	1.4	2.6	2.5	4.1	7.1	×	×	×
c7g.4xlarge	0.4	0.6	1.1	1.4	2.7	2.4	4.4	6.9	×	×	×
c7g.8xlarge	1.0	1.3	1.9	2.5	4.7	4.1	7.8	11.7	12.2	13.5	×
c8g.8xlarge	1.0	1.3	1.6	2.3	4.2	3.8	6.6	10.5	10.6	11.5	×

OpenELM-450M on g5.xlarge

We tested 18 different quantizations for this model & instance. You can see them below, with the most cost-effective option labeled **best**.



Once we start organizing benchmarks with a database, we can’t imagine going on without it.

What questions will *you* ask a database capable of remembering millions benchmarks?