
VENTURI
AI. WITHOUT
THE FLUFF.
VENTURITS.AI

Signup while we wait @ sizer.venturits.ai

Access Code:
ATSL-GLH7

Requires business domain email
Login lasts for 7 days
Export to PDF
Save scenarios

OR

Event Code:
LHPB-XCZ3

3-hour time-limit
No export to PDF
Nothing saved



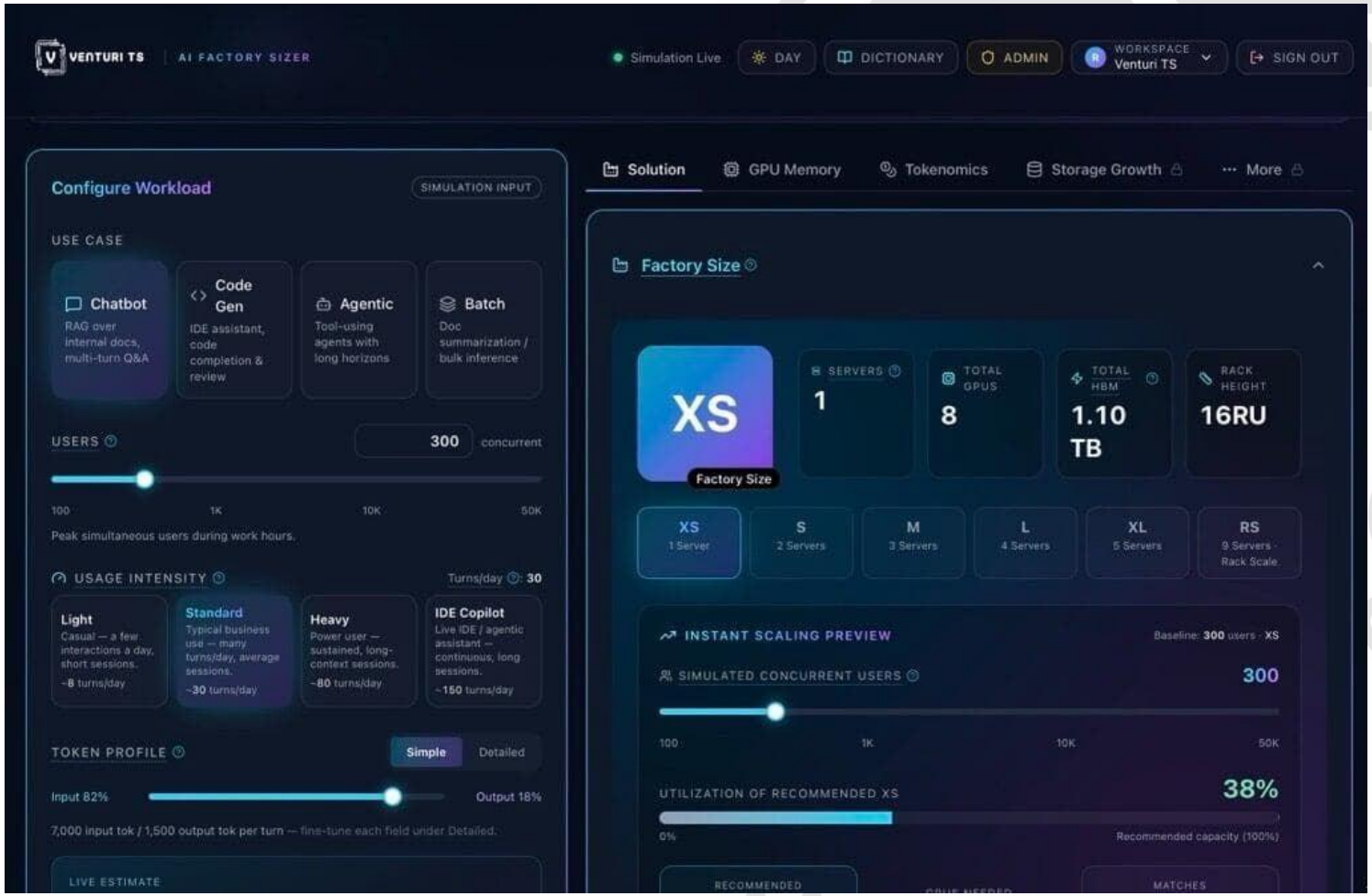
The Economics of AI

Tokenomics:

From One Prompt to an AI Factory

AI TECH TALK 03
MEASURING AI | OCT 16

Factory = the function ≠ the size



Where We Ended Last Time

- Tokens are the smallest unit of measurement
- A prompt becomes context, prefill, decode, and a response
- A simple prompt (100 tokens) explodes to thousands of tokens
- Context, concurrency, latency and model choice shape the demand
- A GPU server runs a model; and an AI Factory delivers a service

Today's Destination

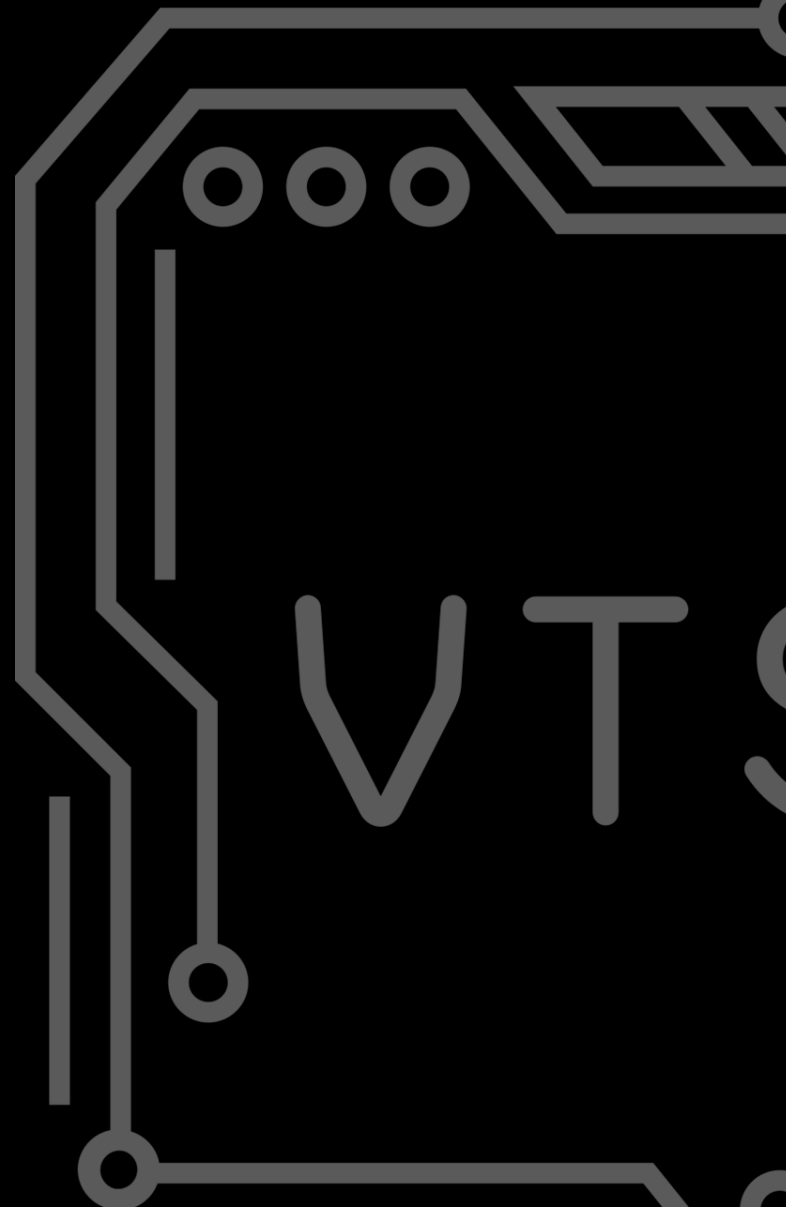
PROMPT | TOKEN | SCALE | FACTORY BOM | COST / TOKEN | SIZER

Small costs become enterprise economics when usage multiplies

0:05–0:13

Five-minute recap

What is a prompt? What is a token?



From Prompt to Token Cost

USER PROMPT

Question, instruction, workflow

INPUT TOKENS

Prompt + context + history

MODEL WORK

Prefill + decode

OUTPUT TOKENS

Generated response

ECONOMICS

Volume × rate or owned cost

The Multiplier is the Story

ONE INTERACTION

A few thousand tokens

PRODUCTION WORKFLOW

More context + more calls

CONCURRENT ADOPTION

More users at the same time

ENTERPRISE SCALE

Persistent, material demand

TOKEN VOLUME × USAGE SCALE × MODEL ECONOMICS = TOTAL COST

0:13–0:21

What does intelligence cost?

High-level components behind every produced token.



“...you essentially pay for intelligence twice, once with money, and again with something even more valuable: the proprietary knowledge you must reveal to make that intelligence useful.”

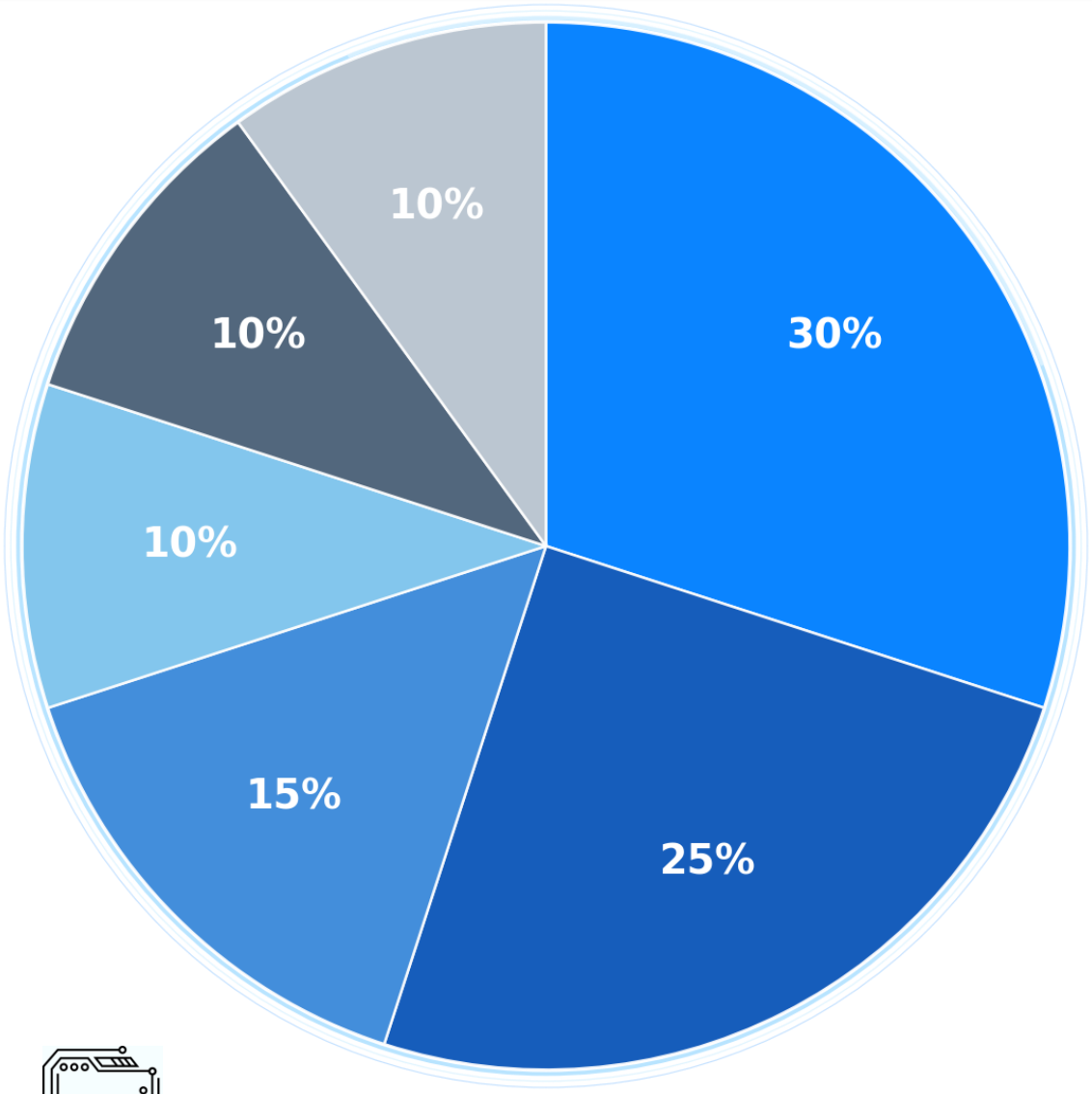
–Microsoft CEO | Satya Nadella

The Intelligence (Factory) Bill of Materials

SERVERS	GPUs	NETWORKING	KV CACHE	OPERATIONS
CPU, memory, storage, chassis	Inference capacity and memory	Move data and synchronize work	Retain useful inference state	Power, cooling, software, support
~$200K (each)	~$30–60K (each)	~$200K (48 ports)	~$500K (200TB)	~$1M (Rack Scale)

The GPU is the visible cost.
The factory determines whether that capacity is productive.

Factory By Percentages



30%
GPU Compute

25%
NAND

15%
Software & Runtime

10%
Storage

10%
Networking

10%
Power & Cooling

Units of Measurement

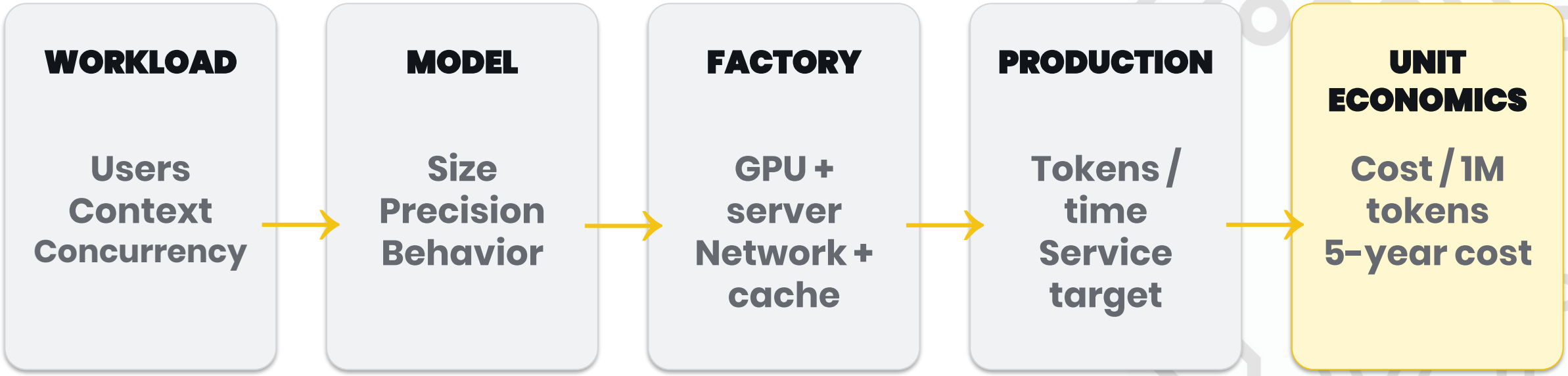
- $/1M Tokens (input vs output vs blended)
- Speed (TTFT) / Latency (time to completion of output)
- Tokens/s/w (efficiency at full load)
- Goodput – factoring in the goodness of the output
- GPU Utilization % (is 100% a good thing?)

Moore's Law & Tokenomics

Two Truths:

1. The latest generation architecture (at scale) will produce the lowest token unit cost – prove it yourself later (>2x the performance, 1.5x the cost)
2. The more density within a unit, the better the "infrastructure gravity escape velocity"

Cost Becomes a Production Model



Change One Variable. Recalculate the System.

MODEL

Memory + throughput

GPU

Fit + performance

CONCURRENCY

Active demand

CONTEXT

Prefill + state

CACHE REUSE

Repeated work

LATENCY

Required headroom

Independent inputs. Interdependent economics.

Why Venturi Created the Sizer

START WITH THE WORKLOAD

Business objectives, users, context, growth

MAKE ASSUMPTIONS EXPLICIT

Known, estimated, and unknown inputs

TEST SENSITIVITY

Observe what changes demand and economics

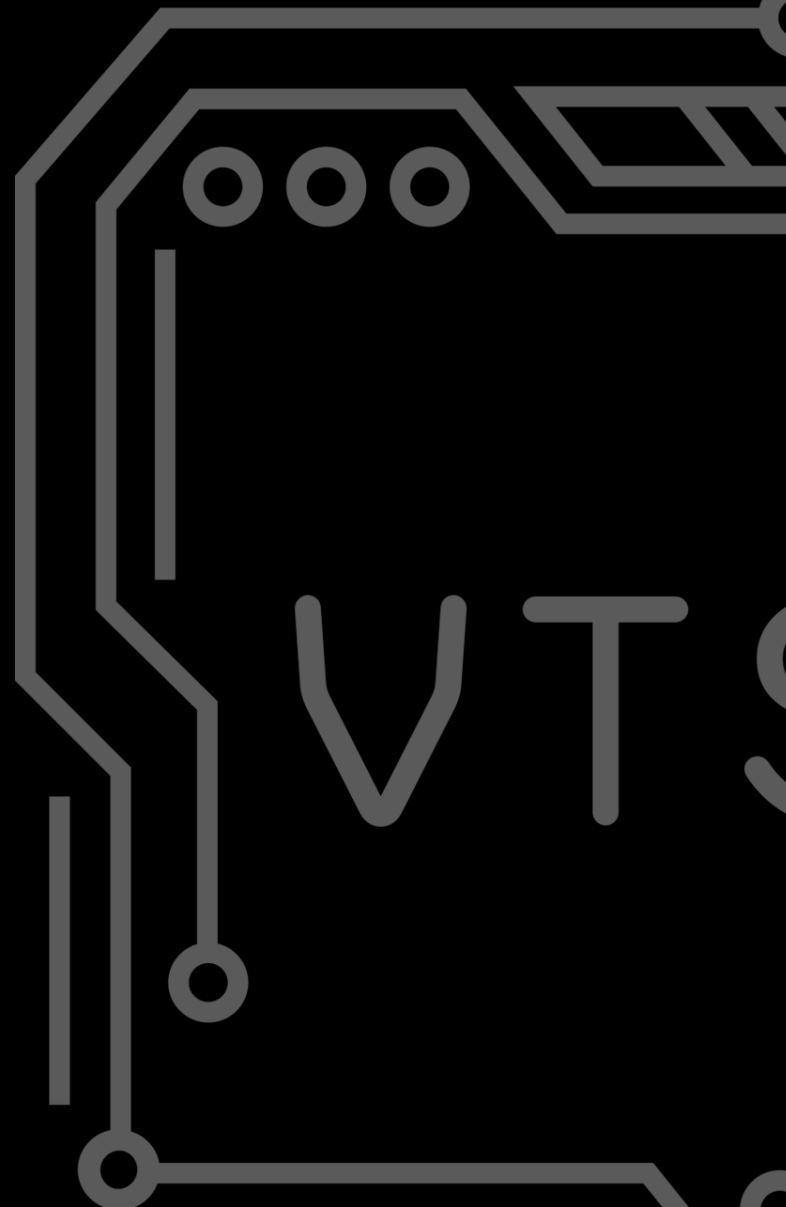
CONNECT THE CONVERSATION

Business requirements to architecture decisions

0:21–0:31

Model the economics

Different examples. Different answers



Consumption vs. Capability

CONSUME

Fast access
Variable expense
Frontier choice

RENT

Dedicated capacity
Flexible commitment
Operational simplicity

OWN

Control
Utilization upside
Strategic asset

There is no universal winner.
The workload determines the economic answer.

A FAMILIAR EXAMPLE

Renting the Building Isn't the Problem.

The risk isn't renting the building. It's leaving your advantage inside.

The question is what you control, and what goes with you when the lease ends.



RENT THE BUILDING
KEEP THE CAPABILITY

You rent the space, but you own what makes the restaurant successful.

- **YOUR RECIPES**
  Your proprietary recipes.
- **YOUR TRAINED TEAM**
  Your trained team.
- **YOUR PROCESSES**
  Your operating procedures.
- **CUSTOMER RELATIONSHIPS**
  Your customer relationships.
- **LEARNING & OPTIMIZATIONS**
  Your experience and improvements.

YOU CAN MOVE.
THE BUSINESS COMES WITH YOU.



RENT THE CAPABILITY
LOSE WHAT YOU'VE BUILT

You get the space, but the landlord keeps what made it successful.

- **YOUR RECIPES**
  Remain with the landlord.
- **YOUR TRAINED TEAM**
  Remains with the landlord.
- **YOUR PROCESSES**
  Remain with the landlord.
- **CUSTOMER RELATIONSHIPS**
  Known by the landlord.
- **LEARNING & OPTIMIZATIONS**
  Remain with the landlord.

YOU MAY HAVE TO START OVER.
THE CAPABILITY STAYS BEHIND.

Compute is the building.

Intelligence is increasingly the recipes, processes, experience, and institutional knowledge inside it.

INTELLIGENCE IS YOUR COMPETITIVE ADVANTAGE.

ARCHITECT AI FOR WHAT'S NEXT.

The Token is the Unit. The Workload is the Economics.

01 A prompt creates input and output tokens

02 Enterprise usage multiplies token demand

03 The factory BOM determines owned production cost

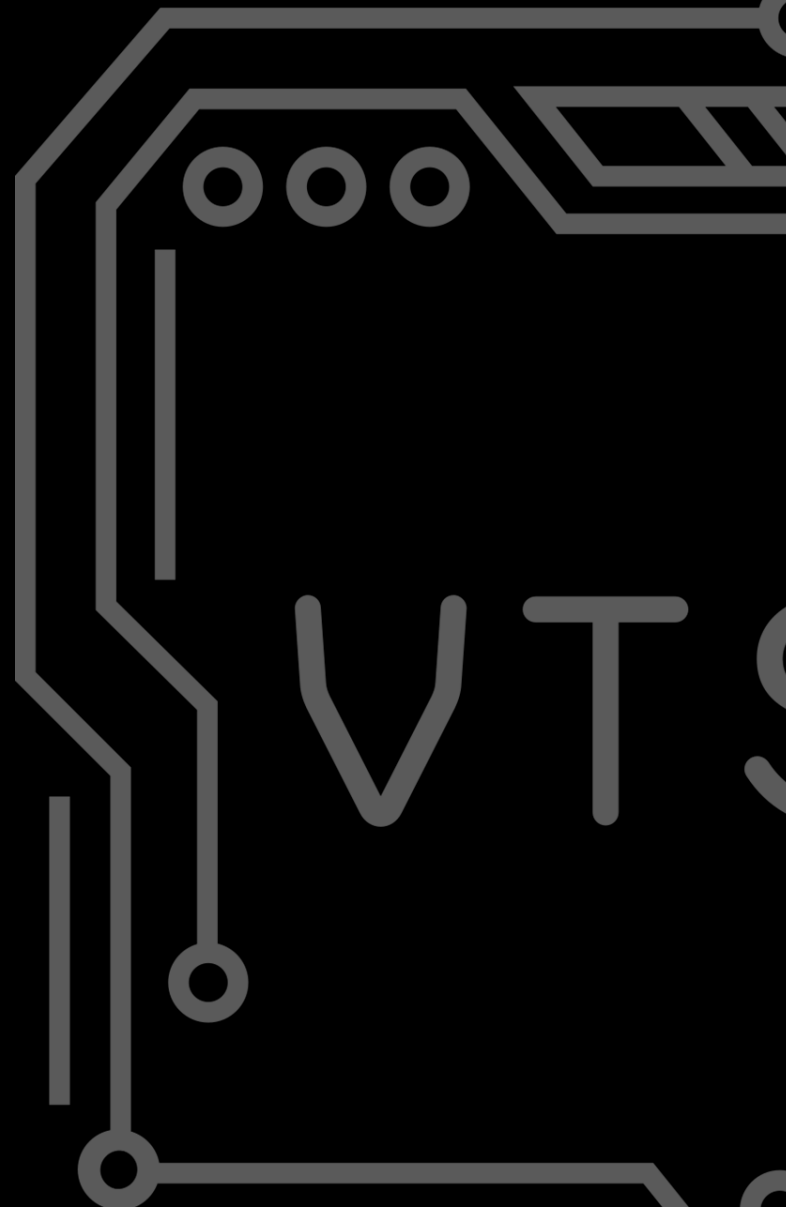
04 Small input changes can alter the whole economic model

05 The Sizer makes assumptions and sensitivity visible

0:31–0:43

Live – Venturi AI Factory Sizer

Real Workloads. From 1 user to 100,000+



The Sizer Roadmap

NOW

- ☑ Workload Planning
- ☑ Architecture Modeling
- ☑ Storage Analysis & Growth
- ☑ Tokenomics Analysis
- ☑ Executive Summary Generation
- ☑ Import Frontier Models Usage
- ☑ Live Infrastructure Scenario Analysis
- ☑ Current GPUs and Models Supported
- ☑ Industry-Specific Use-Cases
- ☑ Efficiency Optimization

SOON

Validation & Benchmark Runbook
Operational feedback & calibration



Thank you.

Learn more @ **venturits.ai**

