INVESTORY THESIS

Every device becomes an inference endpoint.

A 20-year capital allocation framework for energy, chips, memory, data, endpoints, applications, and the trusted execution layer around model-mediated work.

CORE POINT

Inference becomes a metered utility for cognition. The winners own scarce inputs, endpoint distribution, or runtime trust.

I. THE SHIFT

The internet connected documents, people, and services. The next network connects intelligence to every endpoint.

Every device with a sensor, screen, motor, workflow, account, or API becomes attached to inference. The unit of demand is no longer a page view, query, or seat. It is an inference event.

The 20-year opportunity is not only bigger models. It is the full industrial stack required to make model-mediated action cheap, trusted, local when needed, cloud-scale when necessary, and economically measurable.

Inference is an energy-to-application economy.

The stack starts with power, grid interconnects, substations, transformers, land, cooling, and construction. It passes through accelerators, memory, packaging, optics, and data centers. It ends in devices, applications, agents, security, identity, routing, memory, evaluation, and settlement.

The mistake is to analyze AI as one market. It is a chain of bottlenecks. Capital migrates as constraints move: first power and chips, then memory and networking, then endpoint distribution, then data rights and trusted execution.

The 1000x candidate is a default layer, not a single model or one generation of compute.

Mature AI infrastructure can still compound, but venture-scale asymmetry requires a company born where the market is unresolved. The most interesting layer is the neutral endpoint network that routes, secures, meters, evaluates, and governs inference across devices, agents, models, clouds, and data.

The company that wins this layer becomes the operating surface for model-mediated work: which model runs, where it runs, what context it receives, what identity it acts under, what it can spend, what action is allowed, and whether the outcome created value.

IV. MARKET SIGNALS

Endpoint attach is the number to watch.

AI PHONES

912M

Projected annual GenAI smartphone shipments by 2028.

AI PCS

166M

Projected annual AI PC shipments by 2028.

IOT

37.4B

Projected connected IoT endpoints by 2030.

INFERENCE COST

>90%

Expected cost reduction by 2030 for large-model inference.

DATA CENTER POWER

945 TWh

Projected global data center electricity demand by 2030.

CLOUD CAPEX

$830B

Estimated 2026 capex for the top nine cloud service providers.

V. FULL STACK

The opportunity spans the whole machine.

01 Energy and grid

02 Data centers and cooling

03 Chips and packaging

04 Memory and storage

05 Networking and optics

06 Data rights and context

07 Models and inference providers

08 Applications and workflows

09 Security and governance

VI. CAPITAL RULES

1000x capital requires a software-like monopoly on top of a physical supercycle.

01 Own a scarce physical input, an endpoint distribution point, or a trusted execution layer.

02 Avoid undifferentiated GPU rental unless there is a power, financing, utilization, or software edge.

03 Prefer model-independent infrastructure that benefits from plurality across models and clouds.

04 Look for runtime control: identity, routing, policy, memory, evaluation, metering, and settlement before action.

05 Treat proprietary action data as the next scarce dataset: what agents tried, what failed, what worked, and what value followed.