The Risk Transfer Layer for the AI Economy.

Why the compute economy needs insurance and derivatives, and how Forward Compute is building them.

AUTHOR
Forward Compute, Inc.

COVERAGE
Insurance · Swaps · Forwards

SETTLEMENT
Proprietary and white-label index family

KEY TAKEAWAYS

01 Compute is becoming the defining commodity of the AI industrial cycle. The financial architecture and risk-transfer layer around it have not been built.

02 Every prior industrial cycle built its risk-transfer layer alongside its physical infrastructure. The compute economy has not.

03 Forward Compute is building this missing layer: insurance products at Lloyd's, ISDA-documented compute swaps, and assignable forwards. All settling against a purpose-built index family.

EXECUTIVE SUMMARY

A commodity without a risk market.

Compute is becoming the defining industrial commodity of the decade. Goldman Sachs forecasts the AI Infrastructure-as-a-Service market will reach $580 billion by 2030, and the broader software economy it enables is estimated to unlock $2 trillion of incremental revenue[1].

Yet the financial architecture around compute has not been built. There are no deep benchmark prices. There are almost no hedging instruments to address price volatility. There is no mature insurance market for the specific operational, credit, and residual-value risks that define the GPU economy. The risks live entirely on the balance sheets of operators, lenders, and buyers; and as the industry scales toward trillions of dollars in capital deployment, that posture is no longer tenable.

Data centers are built on multi-year horizons, GPU clouds are financed on fragile revenue profiles and a volatile asset base, and buyers commit to contracts they cannot hedge. Every previous commodity cycle - oil, grain, electricity, freight - developed this infrastructure. Compute has not.

FOUR EXPOSURES, CURRENTLY UNHEDGED

01 Price volatility and dispersion on compute pricing

02 Operational risk from outages and performance breaches

03 Residual value risk on GPU inventory

04 Credit and receivable risk on reserved instances and forward contracts

We address these exposures through three product families: insurance solutions placed with Lloyd's syndicates and rated carriers, compute swaps documented to ISDA standards, and standardized OTC forward contracts. All settle against a purpose-built index suite, engineered to institutional governance standards, in partnership with key data providers.

ROADMAP & FRAMING

Our roadmap begins with bilateral placements and OTC transactions, progresses to standardized contract forms and expanded carrier participation, and culminates in a regulated, cleared exchange for compute derivatives. At each stage, the data generated by our transactions becomes the raw material for the next phase of standardization.

This is a whitepaper about why this architecture is necessary, why now is the moment to build it, and why the institutional, transaction-based approach, rather than a retail-first approach is the only one that can scale.

01 Compute is the commodity of the AI industrial cycle

Since the "ChatGPT moment" in late 2022, artificial intelligence has moved from research curiosity to principal driver of technology investment. It is increasingly plausible that AI will be the main driver of economic growth for the remainder of this decade. If that is correct, then compute - the hardware, software, and physical infrastructure that runs AI models - is the underlying raw material that enables it.

But a commodity is not a commodity until it is standardized, priced, indexable, and tradable. Compute, as of 2026, is none of these things in any rigorous sense. Spot prices for GPU-hour rentals are opaque, low volume and highly dispersed. Long-dated contracts are bilateral, illiquid, and structured around physical delivery with punitive exit terms. There is no regulated venue to trade compute forward, no standardized documentation beyond what each cloud provider unilaterally sets, and no insurance market addressing the specific risks that the commodity creates.

Compare this to other commodities at comparable scale. Oil, natural gas, grain, electricity, carbon. Each has a developed derivatives market, specialized insurance products, and institutionally governed benchmark indices. Those markets did not emerge overnight. They developed in response to the same structural problem the compute economy faces today: operators cannot build at scale without the ability to transfer risk.

Compute has become an input factor at industrial scale. Every industrial input at that scale eventually develops a financial and risk-transfer architecture.

The question is who builds it, and with what design principles.

HIGHLIGHT

> "A commodity is not a commodity until it is standardised, priced, indexable, and tradable."

EXHIBIT 1.1 · SELECTED DERIVATIVE MARKET VOLUMES — NOTIONAL, 2024 AND 2025[2]

MARKET	2024 NOTIONAL	2025 NOTIONAL	SCALE (2025)
Oil - crude & refined (ICE)	~$52T	**~$56T**	
Natural gas (ICE)	~$10.1T	**~$12T**	
Soybeans (CME)	~$4.4T	**~$3.8T**	
Corn (CME)	~$2.3T	**~$2.5T**	
Electricity (EEX, global power)	~$1.1T	**~$1.0T**	
Iron ore (SGX)	~$0.2T	**~$0.2T**	
Freight - dry & tanker (FFAs)	~$0.1T	**~$0.2T**	
Carbon (ICE EUAs)	~$121B	**~$110B**	
Compute (today)	—	—	NO INSTRUMENT

Source: CME Group, ICE, EEX, SGX, Baltic Exchange. Notional turnover, rounded. See detailed methodology in appendix.

02 Historical industrial precedents.

Every major industrial cycle of the last two hundred years has followed the same pattern. A cluster of breakthroughs creates new capabilities. Capital floods in to build the supporting infrastructure. Operators race to scale. And before the operating economics of the new industry are fully understood, a cycle of boom and bust plays out.

The pattern is so consistent that it has its own historiography: the canal mania and railway mania of the nineteenth century; the petroleum capital cycle that reorganised the world energy system; the build-out of long-haul telecoms in the 1990s and the dot-com correction that followed. The infrastructure built in each of these cycles persists. The original capital structures rarely do.

HIGHLIGHT

> "The infrastructure built in each of these cycles persists. The original capital structures rarely do."

01

Cycles of innovation, cycles of capital

The picks-and-shovels layer is where capital concentrates earliest and most aggressively, and where the pain of overshoot is felt most acutely. Many of the firms that laid the railroads went bankrupt before the railroads ever turned a profit. Many of the carriers that built transatlantic fibre were absorbed at cents on the dollar after the dot-com correction.

The end-state was real value creation in every case: networks, refineries, fibre, generation, supply chains. But the route there was littered with destroyed equity. The lesson, repeated through every cycle, is that being right about the long-term value of an infrastructure layer is necessary but not sufficient. An operator also has to survive the volatility on the way to the steady state.

02

Adoption lag and the boom–bust mechanism

The boom-bust pattern is mechanical, not cultural. The diffusion of any new general-purpose technology follows an S-curve: innovators and early adopters first, then early and late majorities, then laggards, on a timeline that is consistently longer than capital allocators assume.

Demand from the early-adopter cohort can be intense, and be easily mistaken by suppliers for the shape of the steady-state market. The capacity expansion that follows is calibrated to that initial misread, and is corrected – sometimes violently – when the broader majority adopts more slowly than projected.

When assumptions change, the entire infrastructure complex gets repriced, and ruin can follow.

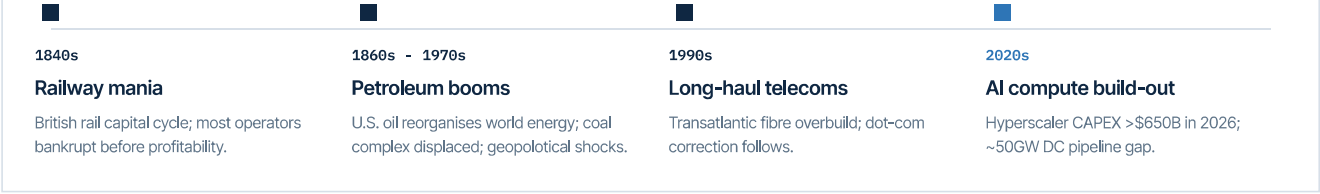
03

The AI cycle now under way

The AI compute build-out fits the pattern. Early demand has come from a concentrated set of hyperscalers and frontier model labs training large language models, and from a wave of application-layer startups building inference-driven products on top of those models. Broader enterprise adoption is real, but it is lagging the deployment curve assumed by today's capital plans.

Signals of recalibration emerged early. In 2025, Microsoft had cancelled several hundred megawatts of US data-centre lease commitments, and Satya Nadella publicly acknowledged the likelihood of an "overbuild" of AI infrastructure even as Microsoft reiterated its $80 billion fiscal-year CAPEX[3]. By April 2026, Bloomberg and Sightline Climate reported that of the roughly 12 GW of US AI data-centre capacity announced for 2026, only about 5 GW was under active construction; the cumulative pipeline gap of announced-but-unbuilt capacity through 2032 approached 50 GW[4].

EXHIBIT 2.1 · INDUSTRIAL CYCLES, CAPITAL DEPLOYED BEFORE RISK MARKETS MATURED



OBSERVATIONS

What operators should expect.

The hyperscalers have not retrenched on CAPEX: Alphabet, Amazon, Meta and Microsoft are still expected to spend over $650 billion on AI infrastructure in 2026. But the conversion of those dollars into energised megawatts and productive GPU racks has consistently slipped. Moreover, it is unclear how fast enterprises are ramping up the conversion of their business model to AI, despite the numerous layoffs announced at recent quarterly earning releases.

These are not yet the conditions of a bust, but they are the conditions in which a bust becomes possible: capital committed faster than it can be deployed, demand assumed to be exponential while broader adoption is still climbing the S-curve, and a small set of large infrastructure providers underwriting the volume on which most operators have built their business plans.

The question for any operator deploying capital into AI compute today is no longer whether the long-term opportunity exists, but how to sustain a balance sheet through the volatility that has bridged every prior cycle to its next steady state.

03 A brief history of risk transfer.

The industrial cycles described above produced more than infrastructure. They also produced the financial instruments that allowed each new economy to function without periodically destroying its operators. Forwards, futures, options, swaps, and the modern insurance market did not arrive as ready-made products. They were built, painfully and iteratively, by commercial actors confronting risks that could ruin them.

The pattern is consistent. A new industry concentrates capital around a small number of operating risks. Early operators absorb those risks on their balance sheets and many fail. Survivors and lenders look for ways to lay off the parts of the exposure that cannot be controlled internally. Out of that demand, hedging instruments emerge: first as bespoke bilateral arrangements, then as standardised contracts, eventually as exchange-traded products with regulated settlement and central clearing.

The same arc plays out across cotton, grain, oil, freight, electricity, and credit. Each new instrument is built first to prevent ruin, not to enable speculation - the speculator is the necessary by-product of a market that exists to serve the hedger.

HIGHLIGHT

> "Each new instrument is built first to prevent ruin, not to enable speculation. The speculator is the necessary by-product of a market that exists to serve the hedger."

01

Forwards as input price stabilizers, not speculation

Forward contracts are among the oldest financial instruments in recorded history. The earliest documented examples date to Mesopotamia. Their purpose was always commercial: give a farmer certainty that the grain would sell at a profit-making price; give the buyer certainty of supply at a known cost. Take volatility off both sides so that they can focus on running their businesses.

The modern framework is identical. Airlines hedge jet fuel so they can price tickets with confidence. Utilities hedge natural gas so they can set retail tariffs. Farmers hedge grain so they can finance the next planting season. In each case, the derivative is a tool for operating confidence, not speculation. A secondary cast – arbitrageurs, speculators, market makers – provides the liquidity that makes these markets function. But the existence of the markets is justified by the hedger's need, not the speculator's interest.

02

Insurance as the other half of the stabilizer set, for operational and balance sheet risks

Derivatives address price risk. They do not, on their own, address operational, counterparty, balance sheet, or asset-depreciation risk. For those, the corresponding instrument is insurance.

Every capital-intensive industry relies on a specialized insurance market. Aviation has hull and liability cover. Shipping has hull, cargo, and P&I. Energy has business interruption, property, and political risk cover. Construction has performance bonds and builders' risk. In each case, operators pay a premium to transfer the tail of specific outcomes to a market – typically Lloyd's, the specialty London and Bermuda markets, the global reinsurance complex and the alternative risk transfer markets – that is purpose-built to absorb that risk.

> **The compute economy will need both derivatives (for price risk) and insurance (for operational and balance sheet risks). These are complementary, not substitute, products. Treating them as a single "risk transfer layer" is the correct conceptual frame.**

A WORKED EXAMPLE · EXHIBIT 3.1

Aviation's response to the geopolitical fuel shock

Jet fuel is typically a third or more of an airline's operating cost, and its price moves on geopolitical timelines that no airline can influence. Russia's invasion of Ukraine in 2022 and the escalation of Middle East conflicts through 2024–2026, including the Iran war that began in early 2026, drove jet fuel prices toward $150 per barrel[5] in spring 2026.

Ryanair — SURVIVED

~80% of fuel needs hedged at ~$67/bbl through March 2027

Blended cost held below $90/bbl. Kept flying summer schedule and planned growth.

Spirit Airlines — LIQUIDATED

0% comparable hedge book in place after two prior bankruptcies

Ceased operations 2 May 2026[6], citing "the sudden and sustained rise in fuel prices."

Ryanair survives the Iran war because it bought protection. Spirit Airlines does not, because it could not. The same dynamic will play out among AI compute operators when the next supply or demand shock arrives, and will separate businesses with a risk-transfer program from the ones without.

04 The specific exposures in the GPU economy.

To build the right instruments, it is essential to be precise about which exposures need to be transferred and who carries them today. The five most important sit on operator and lender balance sheets, almost entirely uncovered.

HIGHLIGHT

"Input-cost moves flow directly into rents on memory-bound GPUs, and they happen in weeks, not in the years assumed by depreciation schedules."

EXHIBIT 4.1 · EXPOSURE REGISTER

REF	EXPOSURE	DESCRIPTION
4.1 PRICE	**Price volatility & dispersion**	GPU pricing deflates over time but the path is non-monotonic. Supply shocks and efficiency breakthroughs move prices in weeks, not in the years assumed by depreciation schedules. As such pricing follows a clear deflationary trend, but a volatile cycle oscillates around the trend. That cycle is why we need derivatives to hedge it.
4.2 OPERATIONAL	**Operational, outage & performance risks**	A GPU cluster outage costs the buyer more than the cluster ever generated in revenue for the provider. Standard service credits do not compensate for business loss.
4.3 RESIDUAL	**Residual value risk**	GPU clouds are depreciating hardware financed with debt. Both inventory and receivable collateral erode rapidly; which can lead to asset-liability duration mismatch and material balance sheet risk.
4.4 CREDIT	**Credit & receivable risks**	Reserved instance contracts generate multi-year receivables from counterparties whose credit quality is opaque. Cash runway depends on future fundraising and product monetization. The entire stack from cloud operator to lenders is at risk.
4.5 PC SPREAD	**Power–compute spread**	The economics of a GPU fleet depend on the spread between compute revenue and electricity cost. As inference grows, compute pricing becomes structurally regional.

EXHIBIT 4.2 · THE FIVE EXPOSURES, IN DETAIL

4.1 Price volatility and dispersion

The consensus view of GPU pricing is that it deflates over time, and this is broadly correct as a long-run trend. Each new architectural generation compresses the economic value of the previous one. The published Silicon Data H100 Rental Index declined approximately 23% between September 2024 and June 2025 alone[7].

But the price path is not monotonic. Supply shocks, such as cluster outages, geopolitical disruption, tariff changes, or simple capacity exhaustion, can drive prices up sharply and without warning. The 2025–2026 cycle of memory shortages is the clearest recent example: contract prices for HBM3E rose approximately 20% for 2026 deliveries, while Samsung's DDR5 32GB module prices were reported up from $149 to $239 (a 60% increase) in November 2025 alone[8]. These input-cost moves flow directly into rents on memory-bound GPUs, and they happen in weeks, not in the years assumed by depreciation schedules.

The demand side moves with at least as much volatility. Through 2024 and 2025, the rapid adoption of agentic coding tools, such as Anthropic's Claude Code, OpenAI's Codex, the open-source project OpenClaw, and a growing population of independent agentic frameworks built on top of them, introduced step-changes in inference demand that no provider could have priced into capacity plans six months ahead. Each new use case that lands at scale shifts the working-hour-equivalent of compute consumed per active developer or knowledge worker by an order of magnitude. The cycle of "new product launches → inference burst → capacity exhaustion → spot price spike" has now repeated several times, and the cadence is accelerating rather than smoothing out.

Efficiency shocks pull the same price in the opposite direction, with significant force. The release of DeepSeek-V3 and DeepSeek-R1 in early 2025 demonstrated that frontier-class reasoning could be trained for far cheaper compute CAPEX than other state of the art AI LLMs, through a combination of FP8 mixed-precision arithmetic, mixture-of-experts architectures, and Multi-head Latent Attention[9]. Alibaba's Qwen series, distilled and refined through 2025–2026, achieved comparable inference quality with materially lower memory and compute

footprints. Each such advance compresses the unit economics of inference: the same task is served at a fraction of the previous GPU-hour cost. That translates directly into downward pressure on rents for the hardware running older or less optimised stacks. The economic value of a GPU is set neither by its silicon nor by its book depreciation schedule; it is set by the most efficient model that the market is currently willing to deploy on it.

The supply chain itself contains multiple structural bottlenecks that translate geopolitical tension into compute price volatility:

- The Spruce Pine mining district in North Carolina supplies 70–90% of the world's high-purity quartz, processed into the fused-silica crucibles used to grow the silicon ingots from which all advanced chips are cut[10].
- TSMC, based in Taiwan, holds the dominant share of leading-edge semiconductor manufacturing. In Q3 2025, processes at 7-nanometre and below accounted for 74% of TSMC's wafer revenue, and the company is the principal foundry for Nvidia, AMD, Apple, Broadcom, and Qualcomm[11].
- ASML in the Netherlands holds a monopoly on the EUV lithography equipment required to fabricate those chips, including the High-NA systems used at 2-nanometre and below[12].
- High-bandwidth memory, essential for every AI cluster, is supplied by only three manufacturers: SK Hynix, Samsung, and Micron. SK Hynix held approximately 57–62% of the HBM market through 2025 by virtue of its first-mover position with Nvidia. As of late 2025 all three suppliers reported HBM capacity sold out through 2026; Samsung's memory chief publicly warned in April 2026 that significant shortages would continue through at least 2027[13].

A commodity with this supply-chain profile and this level of price dispersion is, by definition, one where the ability to hedge has commercial value.

4.2 Operational, outage and performance risks

A GPU cluster outage costs the buyer more than the cluster ever generated in revenue for the provider. Training runs fail and must be restarted, sometimes from early checkpoints. Inference revenue stops outright. Enterprise SLAs break, triggering downstream contractual liabilities. Reputational fallouts impact the entire value chain. The opportunity costs are hard to quantify.

Standard cloud contracts offer service credits capped at a percentage of the monthly bill – typically enough to refund the affected service hours, nowhere near enough to compensate for business loss. The exposure sits with the buyer, uninsured and uncompensated. For a business whose inference product drives millions of dollars of daily revenue, this is a material, structural unhedged exposure.

The product that addresses this is parametric insurance: a policy that pays a predefined amount when objective performance data confirms an outage has occurred, with no loss-adjustment or dispute process. The payouts adapt to the customer's use cases. This is the same instrument used to cover weather in agriculture, earthquakes in property, and flight delays in travel. It is well-suited to the compute context because the triggering data – such as cluster uptime, latency, delivered performance – is already measured continuously and can be made auditor-verifiable.

4.3 Residual value risk

A GPU cloud's balance sheet is primarily composed of depreciating hardware assets financed with debt. Leverage ratios vary across the sector - high-yield neoclouds carry debt-to-enterprise-value ratios well above 60 percent, while investment-grade structures backed by anchor blue-chip customers (CoreWeave's DDTL 4.0 facility, secured by a $19 billion Meta master services agreement, is the first such example) sit lower. In every case the debt is underwritten against two pieces of collateral: the physical inventory (GPUs and servers) and the contracted customer receivables (reserved instances)[14].

Both of these collateral assets depreciate rapidly. The hardware loses economic value as new architectures arrive; the receivables roll off as reserved instance contracts expire. If the operator cannot reinvest capital to refresh the fleet on schedule, the collateral erodes faster than the debt amortises, which triggers covenant pressure, refinancing difficulty, and, in the tail case, default. The depreciation outcome itself is driven by three distinct forces, only one of which is genuinely predictable:

- Technical depreciation. The smooth, calendar-driven decline in economic value as a piece of hardware ages. This component is well-understood, models cleanly, and is the part lenders already price in through book depreciation schedules. Failure rates will also be subject to OEM warranties, usually in-force for 3 years.
- Supply-side innovation. Step-changes in value when a new generation of hardware ships, when a more efficient model architecture changes the unit economics of inference, or when software stacks extract more performance from existing silicon. Discrete, partly anticipated, but consistently faster than book schedules assume.
- Supply and demand shocks. Sharp, unforecastable moves in either direction. The launch of consumer-grade generative AI, the rise of agentic developer tools, and the spread of inference-heavy workloads through 2024–2025 drove demand spikes that lifted realised prices well above trend. The 2025–2026 HBM and DDR5 shortages compressed available supply and pushed memory-bound GPU rents up by similar magnitudes. These shocks dominate variance in realised residual value, and they are exactly what financial markets - not insurance markets - are organised to price.

A residual value guarantee addresses this directly. It is a contract that pays out if the realised market value of a defined GPU inventory falls below an agreed floor at a defined future date. Economically, this is a put option on the GPU price – equivalently, a financial guarantee on

collateral value. The seller receives an upfront premium and assumes the downside between the strike and zero; the buyer keeps all upside above the strike and removes a defined slice of tail risk.

The same instrument has been used for decades in aircraft leasing, auto leasing, and heavy equipment finance, where it can be structured as either an insurance policy (when written by a licensed carrier on indemnity principles) or a derivative (when written by a financial counterparty on parametric or mark-to-market terms). The legal wrapper differs; the economics and exposures are broadly the same.

The wrapper choice has practical consequences. An insurance contract requires an insurable interest, indemnifies actual loss, and benefits from the regulatory and tax treatment afforded to insurance. A derivative settles on observable index or market values without a loss event, can be marked to market daily, and sits on bank or fund balance sheets. Both wrappers can produce the same economic payoff to the operator; which one fits depends on the counterparty offering the cover, the lender's preference for collateral substitution, and the auditor's willingness to recognise the protection in covenant calculations.

Applied to GPU inventory, the result is the same in either form: operators can transfer the depreciation tail to a market – insurance or financial – that prices it independently of the operator's own balance sheet. Lenders see lower expected loss-given-default and can advance more against the same collateral. Operators see longer effective useful-life assumptions and lower cost of capital. The instrument turns a balance-sheet exposure into a hedgeable line item.

4.4 Credit and receivable risks

Reserved instance contracts and forward compute agreements generate multi-year receivables from counterparties whose credit quality is typically opaque. Many of the largest compute buyers are venture-funded AI businesses whose cash runway is a function of future fundraising, product monetization, and model performance – none of which are easily underwritten by a traditional credit analyst.

The exposure sits with the provider and, through the financing structure, with the provider's lenders. When a major customer churns, defaults, or restructures, the cash-flow impact reverberates through the entire stack - from customer, to GPU operator, to DC owner, to providers of capital. Lenders respond with concentration limits that constrain the operator's commercial flexibility.

Credit insurance is the standard response. Trade credit and specialty credit insurers already underwrite similar risks in other sectors and for the hyperscalers complex. They are well-positioned to participate in compute, provided that indices and transaction data exist to support underwriting and pricing.

4.5 The power–compute spread

GPU clusters consume electricity at industrial scale. The economics of a data-center-hosted GPU fleet depend on the spread between compute revenue and electricity cost. In regulated or liberalized power markets, electricity prices can move violently – a dislocation in one market sets the marginal cost for the compute delivered out of that region.

This regional dependency is reinforced from the other side of the equation as the workload mix shifts from training to inference. Training is comparatively portable: a cluster running a multi-week job is largely indifferent to its physical location, tolerates hundreds of milliseconds of latency to a coordinator, and tends to be sited where power is cheapest and capacity is available. Inference is the opposite. Latency and time-to-first-token are first-order product features for any conversational, agentic, or real-time application; users in São Paulo cannot be served acceptably from Northern Virginia, and traders running latency-sensitive models will pay a premium for compute physically close to their venue. As inference grows from a minority to a majority of cycles consumed – a transition already well advanced – the demand side of compute becomes structurally regional, in exactly the way electricity demand is regional.

The implication is that compute pricing will diverge across regions for reasons that are not arbitrageable away. A unit of inference delivered in Frankfurt is not fungible with a unit delivered in Singapore, any more than a megawatt-hour of German power is fungible with a megawatt-hour of Singaporean power. The regional power market sets the cost; the regional inference market sets the revenue; and the spread between them – distinct in every grid – is what the operator actually earns.

As compute pricing becomes indexed and forward-curve data develops, the ability to trade the spread between compute and power in a given region becomes an obvious hedge for both data centre operators and for commodity trading desks. This directly parallels the development of heat-rate and dark/spark spread products in electricity markets over the past two decades – instruments that exist precisely because power and the fuel that generates it are both regionally segmented and are linked by a deterministic conversion. Compute is the third leg of that structure. The conversion is fuzzier than thermal heat-rate, the hardware is younger than turbines, and the regional segmentation is set by physics rather than by transmission constraints. But the economics rhyme closely enough that the same products will work, and the same trading desks will trade them.

05 The Forward Compute framework.

Our product architecture follows the structure of the exposures described above. Three product families cover the risk transfer stack; supported by a suite of vetted third-party and proprietary indices providing the settlement and valuation data.

HIGHLIGHT

"Three product families cover the risk transfer stack, supported by a suite of indices providing the settlement and valuation data."

EXHIBIT 5.1 · FORWARD COMPUTE PRODUCT ARCHITECTURE

5.1 Insurance-linked solutions

LLOYD'S SYNDICATES AND RATED (RE)INSURERS

PROD 1.1

Outage insurance

Parametric cover paying an agreed amount when objective performance data confirms an outage. Triggered against a performance index. No claims adjustment, no causation dispute, settlement within weeks.

PROD 1.2

Contractual liabilities insurance

Covers the service-credit obligations a provider incurs when SLAs are breached. Enables stronger commercial guarantees without placing performance volatility on the provider's balance sheet.

PROD 1.3

GPU residual value guarantee

Guarantees a floor value on defined GPU inventory at a defined future date, settling against the Residual Value index and observed secondary-market data.

PROD 1.4

Credit cover

Covers non-payment on portfolio cloud and/or datacenter receivables. Turns an unrated private receivable into a financeable asset.

5.2 Compute swaps

ISDA-DOCUMENTED, OTC

A fixed-for-floating swap on GPU-hour pricing. One counterparty pays a fixed rate; the other pays the floating value of a reference compute price index. Notional expressed in GPU-hours. Settlement is periodical and purely financial. No physical delivery: the swap can be layered on top of whatever procurement strategy already exists. Documents to ISDA standards and qualifies as a cash-flow hedge under standard accounting, which means CFOs and risk committees can adopt it within the governance frameworks they already use.

5.3 Forward contracts

STANDARDIZED, ASSIGNABLE

Forward contracts involve physical delivery: a defined volume of compute, on a defined infrastructure type, at a defined future date, at a fixed price. Where the compute swap is a financial hedge, the forward is a procurement instrument with hedging properties built in.

The key departure from a traditional reserved instance contract is assignability. A standardized forward can be sold to another counterparty in the forward market at prevailing prices without egress penalties or multi-year lock-in. This is what makes the instrument a forward rather than a long-dated lease: its economic exposure is separable from the operational decision of where and when to run workloads.

06 The size of the AI infrastructure complex.

Estimating the size of any new market is a guessing game. Rather than build a top-down number from assumed financialisation rates and obtainable share, we present three reference points that bound the question. First, the size of the underlying AI infrastructure complex itself. Second, the historic relationship between underlying physical markets and the notional volume of derivatives traded on them. Third, the size of the major derivative markets that already exist for comparable industrial commodities. The reader can apply their own assumptions; we describe the inputs. Detailed methodology of this section is made available on request.

HIGHLIGHT

"The only honest answer to "how large does this become" is: large enough to require the institutional infrastructure that this paper is arguing for."

EXHIBIT 6.1 · UNDERLYING INFRASTRUCTURE SIZE

McKinsey & Company 2025

$6.7T

cumulative global data-centre capex by 2030

Of which $5.2T is AI-specific.

Bain & Company 2025

$500B

annualised AI capex run-rate by 2030

To meet ~200GW of compute demand.

Goldman Sachs Research 2024

$580B

AI Infrastructure-as-a-Service market by 2030

Within a $2T total cloud market.

> **The ranges are wide, the assumptions differ, and forecasts will be revised. The order of magnitude is the point: AI infrastructure is on track to become one of the largest physical-asset markets in the global economy.**

DERIVATIVE MARKETS TRADE AT MULTIPLES OF THE UNDERLYING NOTIONAL TRADED

Across mature markets, the notional volume of derivatives traded each year is a multiple of the value of the underlying physical or financial market. The multiple is a function of market depth, transaction sizes and frequency, participant diversity, the existence of speculators and arbitrageurs alongside hedgers, and the maturity of the supporting infrastructure. It is not a constant, and it grows over time as a market develops.

Applied as a thought experiment to compute, even a low single-digit multiple of the underlying AI infrastructure spend implies derivative notional volumes that quickly reach the scale of mature commodity markets. The exact multiple compute will reach is a question for the market itself; the existence of a multiple is the point. Markets at this scale do not run without one.

Where these multiples sit relative to physical underlying differs sharply by market. Mature commodity derivative markets trade at 12x to 24x annual production value - corn at roughly 12x, oil at 19x, soybeans at 24x. Interest-rate derivatives sit at 3x to 5x of total debt outstanding. Electricity derivatives, by contrast, remain at less than 1x of physical generation value, reflecting the regional fragmentation and partial financialisation of power markets. Compute will likely follow the electricity arc more closely than the oil arc in its early decades. These are the reference points to keep in mind when estimating where compute derivatives end up. The only honest answer to "how large does this become" is: large enough to require the institutional infrastructure that this paper is arguing for.

INSURANCE MARKETS UNLOCK THE ENTIRE INDUSTRIAL COMPLEX

The insurance side of the risk-transfer layer is additive to the derivative side and follows a different logic. Insurance volumes scale with the value of insurable assets and the perceived severity and frequency of covered events, not with derivative turnover. The defining feature of every mature specialty insurance market is that premium pools are small in absolute size: typically 0.1 to 1 percent of the insured capital base. Yet the existence of cover is what allows that capital to be financed at all. The comparators are consistent across very different asset classes.

Global marine insurance premiums totalled $39.9 billion in 2024 against a commercial fleet valued at over $1 trillion and several trillion dollars of cargo in transit at any moment. Global aviation insurance premiums sit at approximately $7-8 billion against a commercial aircraft fleet of comparable size ($1 trillion), with hull cover priced as a percentage of agreed value and treated by lessors as a precondition for any major delivery. Satellite and space insurance is smaller still - $550 million in premiums on roughly 186 launches and 300-plus insured GEO satellites in 2024 - yet it is the cover that has underwritten the debt financing of every major commercial satellite programme for forty years. Cyber insurance, the newest analogue, reached $15.3 billion in 2024 and is forecast to double by 2030; it now sits as a procurement requirement in most enterprise IT contracts and is the model on which institutional buyers expect operational-technology risk to be transferred. Trade credit insurance sits at $12-17 billion of annual premium covering an estimated $2-3 trillion of insured receivables, and is the legal and capital foundation on which the global factoring, supply chain finance, and receivables securitisation markets operate.

GPU-cluster insurance will follow the same template. Premiums will be a fraction of the underlying capital: that is the nature of the product. What matters is what the cover unlocks. Lenders that today refuse to advance against unhedged depreciation, uninsured outage exposure, or unrated cloud receivables will lend against the same assets once those risks have been transferred to the insurance market. The dollar size of the premium pool is not the measure of the opportunity. The measure of the opportunity is the dollar size of the financing it enables.

07 Path to standardization.

> **The market is not fully formed. It will be built in phases, each generating the data, relationships and structures laying the groundwork for it to reach institutional scale and standards.**

EXHIBIT 7.1 · PHASED BUILD-OUT

PHASE 1 — TODAY	PHASE 2 — NEXT	PHASE 3 — END-STATE
Bilateral placements & OTC transactions Forward Compute operates as a broker and intermediary. Insurance products are placed bilaterally with Lloyd's syndicates and reinsurers. Swaps and forwards are structured and negotiated on an OTC basis. Every transaction generates proprietary data that feeds the index methodology and sharpens the product documentation.	**Standardization & expanded participation** As transaction count grows, contract terms converge. Standardized documentation reduces friction and legal cost per trade. Caps, corridors, and portfolio hedges are introduced. Financial market makers, alternative capital institutions and specialty risk providers are onboarded to provide additional liquidity on both sides of the book. Index methodology is tightened as transaction density increases.	**Regulated venue** A fully-integrated suite of regulated entities such as broker-dealers, swap execution facilities, multilateral trading facilities, and clearing houses are developed in core jurisdictions. The OTC infrastructure built in Phases 1 and 2 is extended with clearing, margining, and exchange-grade data distribution. The asset class becomes tradable at institutional scale, with the regulatory barriers to entry that characterize every mature commodity market.

NOTE · INSTITUTIONAL VS RETAIL

Several entrants approach compute as a retail or crypto-native market: tokenized compute, DePIN networks, hourly spot marketplaces. These have technical merit and serve real user needs, but are unlikely to become the foundation of an institutional derivatives market. Regulated financial products require a credible underlying index, IOSCO-grade governance, dollar-denominated volume, and counterparties whose credit can be underwritten. These properties emerge from enterprise, not retail, activity.

A market built on blue-chip counterparties and institutional documentation has a much shorter path to regulatory approval than one built on blockchain-native primitives, however elegant the latter may be.

Our approach is to build the institutional market first and to extend into adjacent structures – including, in due course, tokenized or DeFi-compatible products – once the regulated foundation is in place.

08 Conclusion.

The compute industry is scaling without financial infrastructure and absent a risk-transfer layer. This is a historically familiar condition. Railroads scaled before their financing architecture matured; so did electricity, so did oil, so did telecoms. In each case, the infrastructure that eventually emerged - insurance markets, derivatives exchanges, benchmark indices, specialized lenders - became a core part of how the industry ran.

We are in the early chapters of the equivalent story for compute. The conditions are right here: a large and growing underlying market, deep structural exposures that cannot be ignored, sophisticated counterparties ready to transact, and a regulatory environment that, correctly approached, welcomes the arrival of mature financial products.

Forward Compute is building for this opportunity with a specific design philosophy. Institutional before retail. Insurance and derivatives together, because the exposures require both. Transaction-based indices, not scraped benchmarks. Carrier- and regulator-grade governance from day one. A patient, multi-phase roadmap that turns each transaction into the raw material for the next step of standardization.

HIGHLIGHT

> "The risk-transfer layer for AI compute will be built. You can count on us."

CONCLUSION

The risk-transfer layer for AI compute will be built. You can count on us.

FORWARD COMPUTE, INC.

hello@forwardcompute.ai
forwardcompute.ai

NOTES & SOURCES

01 Goldman Sachs Research, "Cloud revenues poised to reach $2 trillion by 2030 amid AI rollout," August 2024.

02 Derivative volumes calculated as annual contract volume × contract size × average reference price for the year. Sources: CME Group, ICE, EEX, SGX, Baltic Exchange.

03 Bloomberg / Fortune, "Microsoft cancels leases for AI data centers," 24 February 2025.

04 "Nearly half of US AI data centres planned for 2026 are delayed or cancelled," Bloomberg / Sightline Climate, April 2026.

05 Brent crude and jet fuel reference prices, ICE and Platts (May 2026).

06 TIME, "Spirit Airlines Shuts Down Due to Iran War Fuel Crisis," 2 May 2026.

07 Silicon Data H100 Rental Index, public series via Silicon Data and Bloomberg.

08 TrendForce / Chosun Biz, December 2025–January 2026.

09 DeepSeek-AI, "DeepSeek-V3 Technical Report," December 2024.

10 Sibelco, "Spruce Pine: world's leading high-purity quartz operations."

11 TSMC Form 6-K, Q3 2025 filing with the U.S. SEC, October 2025.

12 ASML 2024 annual report.

13 SK Hynix Q1 2026; Samsung Electronics Q1 2026; Counterpoint Research HBM data.

14 CoreWeave DDTL 4.0 facility, March 2026.

15 McKinsey & Company, "The cost of compute: A $7 trillion race to scale data centers," April 2025.

16 Bain & Company, Technology Report 2025, September 2025.

17 CME Group full-year 2025 release, 5 January 2026.

18 Intercontinental Exchange full-year 2025 release, 6 January 2026.

19 European Energy Exchange annual volumes 2025, 13 January 2026.

20 Bank for International Settlements, "OTC derivatives statistics at end-June 2025," December 2025.