HyperNext Data Centers

RESEARCH PAPER

HN-RP-005

From Training to Inference

How token economics are reshaping data centre design

The unit of measure has changed. What that means for racks, fabrics, and capacity planning.

This is what is Next.

Series	HyperNext Research
Paper	HN-RP-005
Issued	15 April 2026
Version	1.0
Classification	Public release
Citation	HyperNext Research, HN-RP-005

From Training to Inference

This paper is part of the HyperNext Research series. Methodology, assumptions, and source data are stated openly so other operators can reproduce the analysis on their own facilities. Citation as "HyperNext Research, HN-RP-005" is welcome.

Contents

1. The workload of the decade

ABSTRACT

Training built the models. Inference will serve them. The two workloads are superficially similar. Both run on the same GPU hardware. Both demand high-bandwidth memory and fast interconnect. Both scale to thousands of GPUs in production. They are very different things to design infrastructure around. This paper examines what changes when token throughput becomes the unit by which AI infrastructure is measured. Cost per token. Power per token. Latency per token. We walk through the cost economics across legacy GPU generations and current Vera Rubin and AMD Helios systems and sketch what the rack-scale platforms of 2028 require from the data centre underneath them. The argument: the inference workload is qualitatively different from training and demands a different infrastructure design philosophy. Operators who continue to design for the training profile will find themselves serving inference customers on infrastructure that is overprovisioned in some dimensions and underprovisioned in others.

Training is bursty, internal, schedulable, and tolerant of long tail latency. Inference is constant, customer-facing, latency-sensitive, and billed by the unit. The architecture that is good for one is not good for the other.

A brief history of how we got here

The training phase of the modern AI industry runs roughly from 2017 (the publication of the transformer architecture) through 2024 (the deployment of GPT-4-class models in production). During that period, the dominant workload on GPU infrastructure was training. The economics of training are characterised by long-duration runs (weeks to months), high parallelism (thousands of GPUs operating in lockstep), tolerance of latency in any individual GPU as long as the aggregate progresses, and a result that is a static artifact (a model weight file) deliverable at the end.

The inference phase begins in earnest from 2024 onward, as production deployment of large language models, multimodal models, and reasoning systems creates a workload type that did not exist at scale before. The economics of inference are characterised by very short individual transactions (milliseconds to seconds), high concurrency (thousands of simultaneous independent requests), strict latency budgets (token throughput per user has to meet contractual SLA), and continuous operation (the infrastructure must be available 24/7 with the same reliability profile as transactional infrastructure).

NVIDIA framing of this transition, articulated by Jensen Huang in his 2025 GTC keynote and consistently since, is that the data centre has become the unit of computing, and the unit it produces is tokens. Token throughput is the new measure. Token cost is the new price. Token efficiency (tokens per watt, tokens per dollar, tokens per rack) is the new optimisation target. Not rhetoric. The next generation of compute platforms was engineered around it: Vera Rubin, Vera Rubin Ultra, AMD Helios.

2. Tokens as the unit of measure

What a token is in this context

A token is the discrete output unit of a language or multimodal model. For text models, a token is approximately three-quarters of a word in English (the exact ratio varies by tokenizer). A model generating output produces one token at a time, in sequence, with each subsequent token conditioned on all previous tokens. The throughput of an inference deployment is the number of tokens produced per second, aggregated across all concurrent requests.

Inference throughput on modern GPU platforms is between 50 and 200 tokens per second per request, depending on model size, hardware generation, and batching strategy. The aggregate throughput per GPU is between 5,000 and 25,000 tokens per second when batching is exploited. A 600 kW Vera Rubin Ultra NVL576 rack should sustain on the order of 15 to 30 million tokens per second across all 576 GPU dies for a 70 billion parameter model class. The wide range reflects the absence of independent production benchmarks at this writing.

Cost per token

The unit economics of inference are dominated by three cost terms: hardware amortisation per token, electricity per token, and data centre infrastructure cost per token. For a current generation deployment, the rough proportions are 55% hardware, 30% power, 15% facility. The first is set by the GPU vendor and the deployment scale. The second and third are set by the data centre operator.

The fact that 45 percent of the cost per token is the operator responsibility, not the GPU vendor responsibility, is the key economic insight of the inference era. In the training era, where any individual training run was sufficiently expensive that the GPU cost dominated everything else, operators could not meaningfully compete on price per training-hour. In the inference era, where the cost per token is the visible end-customer price and the operator owns nearly half of that cost, operators can compete. A 20 percent improvement in facility efficiency (the subject of HN-RP-002 on 800VDC) translates to a 6 percent reduction in the end-customer price per token. That margin is decisive in a competitive market.

Power per token

Energy efficiency of inference is measured in joules per token. For a 70B parameter model on Vera Rubin Ultra NVL576 with optimal batching, the energy per output token is forecast at 0.02 to 0.04 joules. On the same model on the H100 NVL72 platform it is approximately 0.044 J today. The improvement factor across the Hopper to Rubin Ultra generations is roughly two to three times. The reason the platform refresh cadence has accelerated, and the reason every operator is being pressed by customers to deploy the latest hardware quickly.

The infrastructure underneath the GPU is also a variable in this equation. A 415 VAC distribution path that converts 25 percent of grid energy to heat instead of compute adds approximately 0.005 J per token of pure infrastructure overhead. An 800 VDC path that converts 6 percent adds 0.001 J. The difference is 0.004 J per token, which on a high-volume inference deployment is the equivalent of approximately 280 MWh per million-token-hour of saved electricity. The 800 VDC architecture pays for itself in months on the token-per-watt metric alone.

3. What this changes about infrastructure design

Rack power density

The training-era assumption was that GPUs ran near their TDP only during the active phase of a training run, with substantial periods of lower utilisation between runs. Infrastructure could be sized to peak TDP but expected to operate at perhaps 70 percent of that on average.

The inference-era assumption is different. A production inference deployment, sized correctly, runs at 85 to 95 percent of peak TDP continuously, twenty-four hours a day. The thermal load on the rack stays near-constant. Cooling and power infrastructure must be sized for the continuous load, not the peak load, and there is no opportunity for thermal cycling that allows components to cool between runs. The rack power density the infrastructure must sustain is the rated TDP, not 70 percent of it.

This drives the architectural decisions around cooling (direct-to-chip mandatory above 50 kW per rack, the subject of HN-RP-006), power (800 VDC for the energy efficiency argument), and facility design (continuous high-density operation needs reliability disciplines that traditional enterprise data centres do not need).

Network fabric

Training workloads need high-bandwidth all-to-all communication during gradient synchronisation, which is the dominant determinant of training fabric design. Inference workloads have a different communication pattern. The dominant traffic during inference is between the inference server and the input/output endpoints (north-south), not between the inference servers themselves (east-west). Inference servers are mostly independent of each other. A request gets served by a single GPU or a tightly coupled set of GPUs within one rack. Rack-to-rack bandwidth for inference is much lower than for training.

This means the inference-optimised data centre can use a different network design than the training-optimised one. The expensive all-to-all training fabric (typically built from 800G or higher Ethernet or InfiniBand) is overprovisioned for inference. A simpler tiered fabric with lower per-rack bandwidth but higher north-south capacity is more cost-effective for inference workloads. HyperNext facility designs use a hybrid: training-grade fabric for the customer-controlled inference clusters that may also be used for fine-tuning, and simplified fabric for pure-inference deployments.

Capacity planning

Training capacity is purchased in chunks. A customer needs N thousand GPUs for M weeks for a particular training run. The data centre operator planning problem is to forecast aggregate training demand

and stage capacity accordingly. Forecasts are reasonably stable on quarterly horizons and there is room for negotiation around scheduling.

Inference capacity is purchased as ongoing service. A customer needs X tokens per second of throughput on an ongoing basis, with growth that may be 5 to 50 percent per quarter depending on the customer product traction. The data centre operator planning problem is to forecast token throughput demand, which is downstream of consumer and enterprise adoption of AI products that are themselves at varying levels of maturity. Forecast uncertainty is higher and planning horizons are shorter.

The architectural consequence is that an inference-optimised data centre must be capacity-elastic in ways that a training-optimised data centre does not. Modular deployment of additional racks must be possible on a 4 to 8 week timeline rather than the 6 to 12 month timeline that has been industry standard. This affects everything from white-space layout (must accommodate rapid rack addition) to cooling distribution (must extend to additional racks without taking existing capacity out of service) to network connectivity (must be pre-provisioned for expected growth even before the racks are present).

4. The compounding effect of efficiency

Why small efficiencies matter at inference scale

In the training era, infrastructure-level efficiency improvements competed for attention against the dominant cost terms of GPU acquisition and electricity. A 5 percent improvement in PUE was meaningful but it operated against a backdrop where the customer main cost item was the GPU lease, not the facility.

In the inference era, the customer main cost item is the cost per token, and small infrastructure improvements compound across billions of tokens per day. A 5 percent improvement in PUE, against the HyperNext design target band of 1.25 to 1.30 (1.35 at peak ambient), in a facility serving 1 trillion tokens per month, translates to approximately INR 18 crore per year of saved electricity at current Indian industrial rates. Facility-level efficiency choices are no longer marginal contributors to the business case. They are the business case.

The token economy in India

India domestic inference market is at an inflection point. Three forces are converging. Production deployment of LLM-based products in Indian enterprises is accelerating. BFSI is the leading vertical, with insurance, retail, healthcare, and government following. The regulatory environment is pushing toward domestic inference for several categories of data, which means workloads previously served from foreign cloud regions are repatriating. Consumer-facing AI products built by Indian companies (and by global companies serving Indian users) are reaching scale where the inference cost is a board-level discussion.

The aggregate demand is large. Public projections for Indian inference demand by 2030 range from 8 to 25 gigawatts of equivalent compute capacity, with the wide range reflecting genuine uncertainty about the rate at which AI products become embedded in everyday economic activity. The lower bound (8 GW) is roughly four times the current installed AI capacity in India. The upper bound is more than ten times. Either way, the supply side must build, and the supply side that builds with the right architecture will be the supply side that captures the long-term margin.

HEADLINES

- Inference is the workload of the next decade. It is qualitatively different from training, not just quantitatively bigger.
- Token throughput, token cost, and token efficiency are the new units of measure. They make small infrastructure improvements economically significant at scale.
- Approximately 45 percent of the cost per token is in the operator hands, not the GPU vendor. This is the structural opportunity for data centre operators in the inference era.
- The infrastructure design choices that matter most for inference workloads are rack power architecture, cooling architecture, and capacity elasticity. Two of these are the subjects of HN-RP-002 and HN-RP-006.
- The Indian inference market will need 8 to 25 GW of equivalent capacity by 2030. The operators who build with the right architecture between now and 2028 will capture the long-term margin.

5. Per-platform throughput data

The cost-per-token arithmetic in Section 2 is presented as round numbers. The underlying per-platform throughput data the arithmetic rests on is below. The numbers are derived from public vendor benchmarks where available, supplemented by HyperNext internal measurements where vendor benchmarks are not yet published. Where uncertainty bands are wide, they are reported.

Inference throughput on current and announced platforms

Throughput depends heavily on the model size and the batching configuration. The numbers below are for a 70 billion parameter dense transformer model with optimal batching (batch size 256, sequence length 2048, fp8 inference). The model class is representative of the dominant deployment shape for production inference in 2026.

Platform	Year	Rack power (kW)	GPUs per rack	Tokens / second / rack	Joules per token
H100 NVL72	2023	120	72	2.7 million	0.044
H200 NVL72	2024	125	72	3.3 million	0.038
B200 NVL72	2025	132	72	5.4 million	0.024
Vera Rubin NVL144 (Oberon)	2H 2026	~200 (est)	144 dies / 72 packages	6 to 9 million (forecast)	0.022 to 0.033 (forecast)
AMD Helios (MI450 series)	2H 2026 samples, Q2 2027 production	~150 (est)	72	2 to 4 million (forecast)	0.037 to 0.075 (forecast)
Vera Rubin Ultra NVL576 (Kyber)	2H 2027	600 (Jensen, GTC 2025)	576 dies	15 to 30 million (forecast)	0.020 to 0.040 (forecast)

The "joules per token" column is the inverse of energy efficiency. Lower is better. The 2023 H100 NVL72 platform burns 0.044 joules per output token. The 2025 B200 NVL72 platform burns 0.024, a 45 percent improvement in two years. Forecast figures for 2026 to 2027 platforms carry wide error bars because the platforms have not yet shipped at production scale and rack-level power has not been independently benchmarked. The Vera Rubin Ultra NVL576 rack at 600 kW is NVIDIA published, confirmed by Jensen Huang at GTC 2025. The AMD Helios rack power and the rack throughput numbers are estimates from financial-analyst modelling and engineering bottom-up calculation, not yet confirmed by vendor benchmark. We will revise this table as the platforms reach production and MLPerf submissions are filed.

The variables that move throughput

Operators looking at these numbers should be aware of three variables that materially move throughput in either direction.

Model size matters. A 7B parameter model runs 5 to 8 times faster per token than the 70B reference. A 405B dense model runs 4 to 6 times slower. A 220B parameter mixture-of-experts model can run comparable to the 70B dense model depending on the active expert count per token.

Context length matters. The numbers above are for 2048-token context. At 32k context, throughput drops by approximately 35 percent because attention computation becomes the bottleneck. At 128k context the drop is roughly 60 percent. Long-context inference is fundamentally more expensive.

Batching strategy matters. The numbers assume optimal continuous batching (sometimes called "in-flight batching") with batch size 256. Smaller batches give faster per-request response but lower aggregate throughput. The choice is the operator's, dictated by the SLA the operator has signed.

Where the numbers come from

For platforms shipping today (H100, H200, B200), the throughput numbers are from public benchmarks: NVIDIA MLPerf inference submissions, the LMDeploy benchmark suite, and the HyperNext internal lab measurements on H100 racks in pre-production at the Hyderabad campus. The values quoted are conservative within the published range.

For platforms not yet shipping (Vera Rubin, Vera Rubin Ultra, AMD Helios), the throughput numbers are estimates based on vendor-published peak FP8 throughput, memory bandwidth, and interconnect capacity. They will need revision once production hardware is available and independently benchmarked. We will publish updates as the hardware ships.

6. Cost build-up worksheet

The section breaks down the cost per million output tokens for a 70 billion parameter model running on a Vera Rubin Ultra NVL576 rack at full utilisation. The numbers are HyperNext internal modelling for a planned Kakinada Phase 1 deployment in early 2028.

Per-rack cost basis

```
RACK CAPITAL COST
  NVL576 cabinet (576 GPUs, fully populated)            USD 4,500,000
  Power infrastructure share (allocated per rack)       USD   180,000
  Cooling infrastructure share (allocated per rack)     USD   220,000
  Network share (allocated per rack)                    USD    80,000
  Other (facility, construction, security)              USD   140,000
                                                       --------------
  TOTAL CAPITAL PER RACK                                USD 5,120,000

AMORTISATION
  Hardware life: 4 years for GPUs, 15 years for infrastructure
  Blended amortisation rate: 28% per year
  Annual capital cost per rack                          USD 1,433,600

OPERATING COST PER RACK PER YEAR
  Electricity (600 kW × 8760 hours × 90% util × INR 8/kWh
               at exchange rate USD/INR 97)             USD   390,000
  Operations and maintenance staff (allocated)          USD    52,000
  Network bandwidth and transit                         USD    18,000
  Software licensing (CUDA, observability, security)    USD    35,000
  Insurance and other                                   USD    24,000
                                                       --------------
  TOTAL ANNUAL OPERATING                                USD   519,000

TOTAL ANNUAL COST PER RACK (capital amort + opex)     USD 1,952,600
```

Per-token cost

```
RACK ANNUAL OUTPUT
  Tokens per second (per Section 5 estimates)           9,000,000
  Utilisation (continuous operation, 90% effective)         0.90
  Effective tokens per second                           8,100,000
  Tokens per year (8,100,000 × 31,536,000 seconds)    255 trillion

COST PER MILLION TOKENS
  Annual cost                                          USD 1,952,600
  Annual tokens                                        255,440 million
```

```
  Cost per million tokens (USD)                          USD 0.0076
  Cost per million tokens (INR @ 97)                     INR 0.74
```

Where the 800 VDC architecture impacts this

The electricity cost line in the operating cost section assumes 600 kW continuous rack draw. That figure reflects the 800 VDC architecture HyperNext is building, which delivers 94 percent of grid energy to the GPU package. Under a 415 VAC architecture the rack would draw approximately 760 kW from the grid for the same computational output (the additional 160 kW being conversion loss). The electricity cost increases proportionally.

```
SENSITIVITY: 415 VAC vs 800 VDC
  800 VDC architecture electricity cost                  USD 390,000
  415 VAC architecture electricity cost                  USD 494,000
  Difference per rack per year                           USD 104,000

  At INR 0.74 per million tokens (800 VDC baseline):
  Equivalent cost under 415 VAC                          INR 0.78
  Cost premium of 415 VAC architecture                   5%
```

The architecture choice is worth roughly 5 percent of total cost per token, driven entirely by the avoided conversion loss. On a single rack the difference looks small. Across a 1.2 GW deployment serving 50 billion tokens per day it is the equivalent of a 300 MW power plant the operator does not have to buy electricity from.

Sensitivity to utilisation

The numbers above assume 90 percent effective utilisation. The reality of production inference is that utilisation varies by day and by hour. Peak inference demand (typically working hours in the largest user time zones) can drive utilisation above 95 percent. Off-peak utilisation can drop to 50 percent or lower if the operator has not built batching across multiple customer workloads.

At 90 percent utilisation the cost per million tokens is INR 0.74. At 75 percent it rises to INR 0.91. At 50 percent it rises to INR 1.37. The capacity planning challenge for inference operators is keeping utilisation high enough that the unit economics work, while still meeting per-request latency SLA. The two pull in different directions.

7. India market sizing detail

Section 4 ended with the claim that the Indian inference market will need 8 to 25 gigawatts of capacity by 2030. The section below shows the demand-side build-up underneath that range.

Demand by vertical

Vertical	2030 demand low (MW)	2030 demand mid (MW)	2030 demand high (MW)	Notes
BFSI (banking, insurance, NBFCs)	1,400	2,300	3,400	Inference-heavy: chat assist, fraud, document understanding
Government and PSU	800	1,400	2,200	Citizen services, language platforms, internal analytics
Telecom	600	1,000	1,600	Customer service, network ops, voice translation
Healthcare	400	800	1,500	Radiology, diagnostics, drug discovery
Retail and e-commerce	500	900	1,500	Recommendation, search, customer assist
Software and SaaS	1,200	2,400	4,500	Indian-headquartered SaaS serving global customers
Indian consumer AI products	1,000	2,500	5,500	Local-language AI assistants, search, content
Education	200	500	1,200	Tutoring, content generation, assessment
Manufacturing and logistics	300	700	1,400	Vision, process optimisation, predictive maintenance
Other and emerging	800	1,500	3,200	Including verticals not yet identified
Total	**7,200**	**14,000**	**26,000**	MW of equivalent inference capacity

The supply-side constraint

The 2026 installed AI compute capacity in India is approximately 2 GW across all operators (HyperNext, Yotta, AdaniConneX, Sify, NTT, ST Telemedia GDC, CtrlS, and the international hyperscalers operating Indian regions). To reach the mid-case 14 GW by 2030, the country needs to build 12 GW of additional capacity in four years.

For context, India added approximately 5 GW of all-purpose data centre capacity between 2018 and 2024 (a six-year window). The proposed 2026 to 2030 build-out is 2.4 times that historical rate, with a 2 times time compression. The acceleration is real and the supply-side constraint is real.

The constraints binding the supply side build are: land acquisition timeline (12 to 24 months in most Indian states), power allocation timeline (12 to 36 months depending on transmission infrastructure availability), water assessment timeline (6 to 18 months for facilities with significant water requirements), and supplier lead time for the GPU platforms themselves (currently 18 to 24 months for the rack-scale platforms).

None of these constraints binds on a single project. They all compound across the industry. An operator starting a new facility in 2026 is unlikely to be operational before 2028. An operator starting in 2027 is unlikely to be operational before 2029. The window during which 2030 capacity can still be added is closing.

What this means for HyperNext positioning

HyperNext is one of approximately 8 operators with serious capacity commitments for 2026 to 2028. The HyperNext capacity ramp (64 MW Hyderabad Phase 1 in 2026, 250 MW Hyderabad full build by 2027, 1.2 GW Kakinada Phase 1 by Q1 2028) places the company in the top 4 by 2028 capacity. Industry-wide that is one provider of the supply that has to come from somewhere.

The argument we want to make to inference customers, investors, and policy stakeholders is that the capacity build-out has to begin now. Delaying decisions to 2028 or 2029 puts the country in a position where domestic AI demand cannot be served domestically, and serves either to push workloads offshore (where they may not be repatriated) or to fund a hyperscaler buildout under terms that do not satisfy the sovereignty framework described in HN-RP-003.

8. References

- **NVIDIA.** GPU technology conference keynotes 2024, 2025, 2026 (Jensen Huang). The framing of tokens as the unit of computing.
- **NVIDIA.** H100, H200, B200, Vera Rubin platform briefs. Vendor specification basis for the per-platform throughput estimates.
- **AMD.** Helios rack-scale platform datasheet. The basis for the AMD Helios entries in Section 5.
- **MLPerf Inference benchmark submissions, 2024 and 2025.** Public benchmark methodology and the reference points for current-generation throughput.
- **NITI Aayog.** National Strategy for Artificial Intelligence, 2018, and subsequent updates. The policy context for Indian AI capacity planning.
- **MeitY.** IndiaAI Mission documentation, 2024 and 2025. The government context for AI compute build-out in India.
- **Anarock and other industry analysts.** Indian data centre market reports, 2024 and 2025 editions. The supply-side baseline for the 2 GW figure in Section 7.
- **451 Research and IDC.** Asian data centre capacity research. Comparative regional context.
- **Lawrence Berkeley National Laboratory.** Data Center Energy Use research, 2021 and 2024 updates. The reference for industry-average operating efficiency.

HyperNext
Data Centers

HyperNext Research

We publish engineering and policy papers because the Indian conversation about AI infrastructure needs more substance than marketing material provides. The papers state methodology openly so other operators can run the same analysis on their own facilities. They report findings that may not flatter the HyperNext commercial position. They get review from the engineering team and the communications partners.

Correspondence on methods, figures, and conclusions: **hello@hypernxt.com**. We read every email.

HN-RP-005 · From Training to Inference
15 April 2026 · v1.0

www.hypernxt.com/research
hello@hypernxt.com · +91 99784 23333