

42. Compute Abundance, Infrastructure Scarcity

Status: Public report

Date: 30 August 2026

Author: Njål Gaute Solland

Executive summary

AI is moving toward an unusual economic state: useful compute can become cheaper and more capable at the same time that the physical infrastructure required to deliver it becomes harder to secure.

The apparent contradiction disappears once the stack is separated.

silicon / software efficiency → faster improvement
power / grid / land / cooling / build → slower expansion

The result is **compute abundance inside infrastructure scarcity**.

This matters because the constraint on AI deployment can migrate away from model capability or accelerator price and toward megawatts, substations, interconnection, transformers, cooling systems, construction capacity, financing and time.

The evidence is already visible. The IEA projects global data center electricity consumption to reach around 945 TWh by 2030 in its base case. Lawrence Berkeley National Laboratory's 2025 update estimates US data centers could account for 11.8% of total US electricity consumption by 2030, with a scenario range of 9.5% to 15.3%. JLL expects global data center capacity to almost double from 103 GW in 2025 to 200 GW in 2030. McKinsey's July 2026 US analysis projects roughly 121 GW of data center IT demand by 2030 and concludes that the greater near-term risk may be underbuilding rather than overbuilding.

1. Two different supply curves

"Compute" is often treated as one resource. It is not.

There are at least two supply curves.

Digital supply

- accelerator performance;
- model efficiency;
- quantization;
- inference optimization;
- memory efficiency;
- software scheduling;
- routing;
- utilization;
- algorithmic improvements.

These can improve rapidly and repeatedly.

Physical supply

- electric generation;
- transmission;
- substations;
- interconnection;
- transformers;
- switchgear;
- cooling;
- water or alternative heat rejection;
- permitted land;
- fiber;
- construction labor;
- debt and project finance.

These move on infrastructure timescales.

The IEA explicitly notes the mismatch: a data center can become operational in roughly two to three years, while broader energy infrastructure often requires longer planning and construction lead times.

2. The demand curve is moving faster than the grid

The IEA's base case projects data center electricity consumption to reach roughly **945 TWh in 2030**, more than double current levels and growing around 15% annually from 2024 to 2030.

Accelerated servers—mainly driven by AI—are projected to grow electricity use around 30% annually in the base case.

In the United States, LBNL's 2025 update estimates that data centers could account for **11.8% of total US electricity consumption by 2030**, with a range from 9.5% to 15.3%.

These are not simply estimates of more servers. They imply a large reallocation of electricity growth toward one sector.

3. Capacity expansion is enormous and still constrained

JLL estimates global installed data center capacity at roughly **103 GW in 2025**, rising to **200 GW by 2030**.

That implies nearly 100 GW of new capacity in five years.

JLL estimates this buildout could require up to **\$3 trillion** in combined real estate and tenant infrastructure investment, including roughly **\$870 billion of new debt financing**.

In North America alone, JLL reported more than **66 GW under construction** in mid-2026.

Despite that scale, vacancy remains around 1% and most of the pipeline is already committed.

The inference is important:

Massive construction does not prove excess capacity when demand is being absorbed before completion.

4. The scarcity moves upstream

When GPUs are scarce, the bottleneck is obvious.

When GPUs become easier to obtain, the bottleneck moves.

A modern AI cluster requires far more than accelerators:

- power delivered at the right voltage;
- distribution equipment;
- cooling;
- network fabric;
- space;
- fiber routes;
- operational staff;
- backup systems;
- permission to connect to the grid;
- financing.

The scarce unit increasingly becomes not "a GPU," but:

a power-ready, network-ready, financeable megawatt delivered at the right place and time.

That is a different market.

5. Why efficiency does not automatically solve the problem

A common assumption is that more efficient chips reduce infrastructure demand.

At the workload level, they do.

At the market level, the result depends on what users do with the savings.

If efficiency lowers the cost of inference, then:

- more queries become economical;
- reasoning depth can increase;
- more agents can run continuously;
- more products can become personalized;
- latency targets can fall;
- AI can move into workflows that were previously too expensive.

The physical system therefore experiences the net effect of:

```
efficiency gain  
minus  
induced demand
```

If induced demand grows faster, total infrastructure demand rises.

6. Training gives way to inference

The composition of AI demand is also changing.

McKinsey expects inference to overtake training as the dominant AI workload before 2030. Its 2026 workload analysis projects AI inference demand growing much faster than non-AI workloads.

This matters because inference is not a one-time build event.

Training can be concentrated in large frontier clusters. Inference spreads with user activity, software adoption and product integration.

That makes AI demand more persistent and more geographically distributed.

7. The economic consequence

The value chain can separate into two very different businesses.

Commodity-like layers

- raw model access;
- standard inference;
- generic accelerators over time;
- routing;
- basic orchestration.

Scarce infrastructure layers

- power rights;
- interconnection;
- high-density cooling;
- data center shells;
- construction capacity;
- transformer and switchgear availability;
- fiber and network topology;
- financing of large committed assets.

As digital intelligence becomes cheaper, rents can migrate toward the physical bottlenecks that remain difficult to replicate.

8. Geographic consequences

The constraint is local even when AI demand is global.

A country can have abundant generation but poor interconnection.

A region can have land but no transmission.

A site can have grid access but insufficient cooling or fiber.

A developer can have all three but lack equipment or financing.

This produces geographic arbitrage:

- frontier data center markets;
- behind-the-meter generation;
- microgrids;
- storage;
- gas generation;
- nuclear and geothermal interest;
- new fiber corridors;
- regions with faster permitting.

The geography of AI therefore increasingly follows infrastructure optionality rather than software talent alone.

9. What would falsify the thesis

The thesis weakens if one or more of these occurs at scale:

1. electricity demand from data centers materially undershoots IEA/LBNL ranges;
2. interconnection queues clear much faster than expected;
3. delivered power capacity consistently outgrows compute demand;
4. utilization falls despite cheaper compute;
5. efficiency gains reduce total workload demand rather than expanding it;
6. new power-ready capacity produces sustained high vacancy.

The thesis strengthens if:

- vacancy remains very low despite record construction;
- power connection becomes a larger share of project lead time;
- on-site generation and microgrids proliferate;
- inference growth outpaces efficiency improvement;
- power-ready sites command persistent premiums.

10. Conclusion

The AI economy can produce more intelligence per dollar while simultaneously becoming more constrained by physical infrastructure.

That is not a contradiction.

It is a change in the location of scarcity.

old bottleneck: compute hardware
new bottleneck: delivered infrastructure

The strategic question is therefore shifting from:

| Who has the best model?

to:

| Who can secure the physical conditions required to run intelligence at scale?

Sources

- International Energy Agency, Energy and AI — Energy demand from AI: <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>
- International Energy Agency, Energy and AI — Executive summary: <https://www.iea.org/reports/energy-and-ai/executive-summary>

- Lawrence Berkeley National Laboratory, United States Data Center Energy Usage Report: 2025 Update (June 2026): <https://bies.lbl.gov/publications/united-states-data-center-energy-2025>
 - JLL, 2026 Global Data Center Market Outlook: <https://www.jll.com/en-us/insights/market-outlook/data-center-outlook>
 - JLL, North America Data Center Report Midyear 2026: <https://www.jll.com/en-us/insights/market-dynamics/north-america-data-centers>
 - McKinsey, Powering AI: How real is the risk of overbuilding?, 31 July 2026: <https://www.mckinsey.com/industries/electric-power-and-natural-gas/our-insights/powering-ai-how-real-is-the-risk-of-overbuilding>
 - McKinsey, The future of AI workloads, 24 February 2026: <https://www.mckinsey.com/featured-insights/charts/the-future-of-ai-workloads>
-

This report is research, not investment advice.