Regression-Based Network Load Forecasting for Sustainable Digital Infrastructure - stakeholder report

1. Purpose and context

Our network segment connects a data center to the public internet through two upstream providers. Bandwidth on these links is contracted and billed based on high-demand periods (for example, using a 95th percentile rule). Today, we manage capacity largely by experience and coarse monitoring: we buy enough "just in case", accept a safety margin, and react when traffic patterns change.

This approach is safe but not always efficient. Keeping a large amount of bandwidth permanently idle means extra hardware, power, and cooling. At the same time, underestimating demand can lead to congestion, emergency upgrades, and service risk.

The goal of this project is to make our capacity decisions more informed by building a forecasting model that can tell us, hour by hour, whether the next hour is likely to be quiet, normal, or heavy for our upstream links.

2. What the model does

The model looks at:

- recent hourly traffic on our upstream links,
- typical patterns by time of day and day of week,
- simple indicators of how busy our backend hosts and services are,

and uses this to forecast the peak traffic for the next hour on the upstream connection.

In simple terms:

"Given what the last few hours and the same times on previous days and weeks looked like, what is a reasonable expectation for the next hour?"

The model does not inspect any user-level or content data. It works only with aggregated hourly metrics (how many bits per second were sent/received on each uplink, and how busy the hosts were in total).

3. How good it is

To judge whether the model is useful, we compare it to simple baselines, such as:

- assuming "the next hour will look exactly like the last one", or
- assuming "the next hour will look like the same hour yesterday".

Our regression model consistently performs better than these baselines on held-out periods of data. It captures both short-term inertia (what just happened) and regular daily/weekly patterns.

In normal hours, forecast errors are usually moderate and behave in a stable way over time. For rare extreme spikes, the model cannot predict the exact size of the spike, but it generally recognises that a high-load situation is likely. This is expected: very unusual events are hard to predict from history alone and still require human judgement.

The model is therefore good enough to inform decisions and alerts, but not perfect – it should be used as one signal among others.

4. How you can use it in practice

There are three main ways this model can support operations and planning:

1Operational scheduling
Use next-hour forecasts to avoid scheduling bandwidth-intensive jobs (backups, large data transfers, synchronisations) during hours predicted as "high load". This reduces the risk of hitting capacity during already busy periods.

2Early warning and trend monitoring
Track how often forecasts approach a chosen "comfort zone" (for example, 70–80% of link capacity). If forecasts start spending more time near this zone, it is a signal that current tariffs or capacity may need to be reviewed before problems occur.

3Input for future tariff and capacity planning
Although this project only forecasts one hour ahead, the same approach can be extended to longer horizons (days and weeks) and combined into risk indicators for monthly billing periods. This is a first step toward quantitative support for questions like "Do we need to upgrade this link next quarter?" or "Can we safely reduce overprovisioning?".

In all cases, the model is a decision-support tool. Final decisions remain with network and infrastructure teams, who will combine forecasts with their knowledge of planned changes, maintenance, and business context.

5. Sustainability impact

Better forecasting does not automatically make the network "green", but it gives us the information needed to:

•avoid overprovisioning, where we pay for and power significantly more capacity than we realistically need most of the time;

•run closer to the "right-sized" capacity while still respecting safety margins;

- move heavy, non-urgent data transfers into lower-traffic windows, where the electricity mix may be cleaner and the infrastructure is under less stress.

This supports more resource-efficient digital infrastructure and aligns with our broader goals around responsible consumption of IT resources and climate impact.

6. Limitations and guardrails

There are important limits to what this model can do:

- It is trained on about five months of data from a single router. It reflects the behaviour of this specific segment and time period and does not include annual patterns or rare crisis events.

- It works best for "business as usual" conditions. Sudden architectural changes, new products, or major incidents will require retraining and revalidation.

- It cannot guarantee that we will never hit capacity. It can only indicate likelihood and trends, not certainties.

Because of this, any use of the model should follow simple guardrails:

- Treat the model as one input, not the sole decision-maker.

- Keep conservative safety margins when acting on its recommendations.

- Revisit and retrain the model regularly as traffic patterns and infrastructure evolve.

If we follow these principles, the forecasting model can become a valuable part of our toolbox: it will not replace experience, but it can help us see patterns earlier, discuss capacity with data in hand, and gradually move toward more sustainable use of our network resources.