Regression-Based Network Load Forecasting for Sustainable Digital Infrastructure

1. Project context and long-term goal

We operate a data center network segment connected to the public internet via two upstream provider links (BGP). Network capacity is contracted and billed based on high-demand periods (95th percentile of traffic), so not planning planning bandwidth is both expensive and environmentally costly. In the long run, our ambition is to build a workload-aware forecasting pipeline that can support monthly tariff and capacity decisions: for example, estimating the risk of exceeding contracted bandwidth in the next billing period and evaluating whether upgrades or downgrades are justified.

2. Scope of this capstone: step 1 in the pipeline

This capstone focuses on a deliberately narrower, but foundational regression task:
we predict next-hour peak bandwidth demand on the upstream connection (t+1). The target is defined as the maximum of inbound and outbound hourly peak traffic across the two active uplinks, expressed both in bps and as a fraction of total link capacity. This 1-step-ahead forecast is not the final business goal by itself, but it is the core building block for a future system that will aggregate hourly predictions into billing- and policy-relevant metrics (e.g., predicted monthly 95th percentile, risk indicators over future windows).

3. Why this matters for sustainability

Network capacity is not free: overprovisioning and leaving large amounts of bandwidth idle most of the time implies more hardware, more power, and more cooling than necessary. At the same time, underprovisioning and purely reactive scaling can cause service instability and emergency upgrades, which are also resource-inefficient. Better short-term forecasting enables more resource-efficient digital infrastructure by:

- supporting safer operation closer to the “right-sized” capacity instead of permanent overprovisioning;

- scheduling bandwidth-heavy jobs into lower-traffic windows when the electricity mix may also be greener;

- providing early-warning signals when demand patterns shift and start to approach contractual limits.

This directly connects to **SDG 9 (Industry, Innovation and Infrastructure)** and **SDG 12 (Responsible Consumption and Production)** through more efficient use of network resources, and to **SDG 13 (Climate Action)** by helping avoid unnecessary indirect emissions from overscaled infrastructure.

4. Stakeholders and decisions enabled

The primary stakeholders are:

- Network operations / SRE teams, who need to anticipate upcoming peak hours, trigger alerts, and adjust operational schedules (e.g., backups, large data transfers) away from predicted peaks.

- Infrastructure and financial planning, who care about how often demand approaches contracted capacity, whether the current tariff is appropriate, and whether upgrades or renegotiations are justified.

- Sustainability / ESG roles, who need evidence that digital infrastructure is not significantly overprovisioned and that operational practices are aligned with efficiency goals.

In this capstone, the model is used to:

- identify high-risk hours in the near future (next hour) as a prototype for alerting and scheduling;

- explore which factors (time-of-day, weekly patterns, host workload proxies, business growth signals) drive peak demand;

- demonstrate how hourly forecasts can be aggregated into longer-term risk signals that will later support monthly tariff planning.

5. Data and constraints

The project uses an anonymised, hourly-resolution dataset from one router, covering approximately five months (Aug 2025-Jan 2026). The feature groups include:

- per-uplink traffic statistics (min/avg/max) and measurement quality indicators;

- host CPU utilisation signals for three backend hosts;

- monthly business proxies (e.g., new VMs or users, lagged to avoid leakage).

Key constraints are:

- Routing uncertainty: BGP routing decisions between providers are not fully observable, so we focus on forecasting aggregate peak demand rather than traffic split per provider.

- Limited history (~5 months): annual seasonality cannot be captured; the model reflects patterns in the observed period and must be retrained as new data arrives.

- Single-site scope: results are specific to this router and environment and are treated as a pilot rather than a global policy.

66. Initial hypothesis and research questions

Our initial working hypothesis is that next-hour peak bandwidth demand on the upstream link is not random noise, but follows a structured combination of:

– strong temporal patterns (time of day, day of week, weekly seasonality), and
– medium-term workload trends and business growth signals (e.g. number of active VMs, host utilisation patterns),

such that a regularised linear model with carefully chosen lag and calendar features can outperform naive and seasonal baselines in forecasting next-hour peaks.

More concretely, we explore the following questions:

1Predictability:
To what extent can next-hour peak bandwidth be predicted from past traffic alone (using leakage-safe lag features) compared to naive baselines (e.g. "next hour = last hour" or simple seasonal rules)?

2Added value of weekly structure:
Do weekly lag features ("same hour last week") and rolling statistics improve forecast accuracy in a meaningful and stable way, beyond simple daily patterns?

3Role of workload and business proxies:
Do host CPU utilisation and coarse business activity proxies (e.g. monthly growth in VMs) add a measurable signal on top of time-based and traffic-based features, or is most of the predictive power captured by temporal structure alone?

4Interpretability for operations:
Can the chosen regression model provide interpretable relationships between features (e.g. certain hours, days or workload regimes) and predicted peaks, in a way that is actionable for network and sustainability stakeholders?

The capstone does not aim to fully solve long-horizon tariff optimisation; instead, it tests whether a well-structured regression model can provide a robust, interpretable short-horizon forecasting layer that is worth building upon for future monthly forecasting and capacity planning.

7. Success criteria

Within this restricted scope, a good model is defined as one that:

- **outperforms sensible baselines** (naive and seasonal-naive) in RMSE, MAE, and R^2 on walk-forward time splits;

- maintains reasonably **stable performance across time**, not just on one lucky month;

- remains **interpretable**, so that feature effects can be translated into operational and sustainability insights;

-can be realistically reused on **future data** (new months) as part of a larger workflow that aggregates hourly forecasts into monthly 95th-percentile estimates and capacity risk indicators.

88. Ethical and governance considerations

This project works only with **aggregated operational metrics** (hourly uplink traffic, host-level CPU utilisation, coarse business proxies). No user-level or content-level data is included. The dataset is internal and non-public; only the structure and methods are shared in this project, not raw logs or identifiers.

The main governance risk is cost optimisation at the expense of reliability. A forecasting model that highlights "excess" capacity could be misused to justify aggressive downsizing of links. In this capstone, the model is explicitly framed as decision support, not an automatic controller: final capacity and scheduling decisions remain with human operators, who must consider SLAs, resilience targets, and user impact.

The model is trained on ~5 months of data from a single router, so its validity is limited to this environment and period. It does not capture annual seasonality or behaviour under extreme or crisis conditions. Responsible use therefore requires transparency about:

•the training window and data coverage,

•expected uncertainty and failure modes (e.g. unusual peaks),

•the need for periodic retraining as traffic patterns or routing policies evolve.

Finally, sustainability is treated as a joint constraint with reliability, not a competing goal. Any reduction in overprovisioning or closer operation to capacity should be checked against resilience requirements and remain reversible if performance or user experience deteriorate.

9. Data overview and feature groups

The modelling dataset is built from internal Zabbix exports for a single edge router connected to two upstream providers. After preprocessing and quality filtering, it covers roughly five months of hourly observations (Aug 2025–Jan 2026). Each row corresponds to one UTC hour and includes:

•Time-based features
– hour of day (both as numeric and sine/cosine encoding),
– day of week and weekend/weekday flags,
– a simple time index capturing gradual trend over the period.

•Traffic-based features
– hourly max/avg/min for each active uplink (in and out),
– daily lag features (e.g. `lag_1`, `lag_24`, `lag_48`) to capture short-term inertia,
– rolling statistics over recent windows (e.g. mean and std over the last 24 hours),
– a weekly lag (`lag_168`) in the extended feature set to capture "same hour last week" patterns.

•Workload and business proxies
– host-level CPU utilisation aggregates (idle/busy time converted into "busy" features) for backend hosts,
– coarse monthly VM growth indicators, lagged to avoid target leakage and used as a proxy for underlying demand growth.

•Data quality indicators
– flags for low-quality or interpolated hours,
– simple counters describing how many bad-quality measurements are present in a rolling window.

The target for this capstone is the next-hour peak load on the upstream connection, defined as the maximum outbound traffic across active uplinks (expressed both in bps and as a fraction of nominal link capacity). Full exploratory analysis, including justification for each feature block and the handling of missing or low-quality data, is documented in `EDA_and_FE.ipynb`.

10. Project files and navigation

The project is organised as a small, self-contained folder with separate notebooks for exploration and modelling:

•`EDA_and_FE.ipynb`
Exploratory Data Analysis and Feature Engineering.
This notebook documents the data source, cleaning steps, traffic and seasonality patterns, peak analysis, and the rationale behind the final feature set.

•`regression_models.ipynb`
Model development and evaluation.
This notebook implements baselines, trains and compares regression models (Linear, Ridge, ElasticNet, XGBoost), performs diagnostics, and selects the final Ridge model with daily and weekly lag features.

•`data/`
Contains the final clean modelling dataset (`network_hourly_clean.csv`).

•`reports/README.md` (this document)
Provides the high-level problem framing, sustainability context, ethical considerations, and a roadmap to the notebooks and assessment criteria.

11. Modelling approach and validation strategy

We focus on a compact, well-justified set of models suitable for tabular time-series features:

•two baselines:
– Naive persistence: ŷ[t] = y[t–1] ("the next hour is like the last one");
– Seasonal naive: ŷ[t] = y[t–24] ("the same hour yesterday");

•regularised linear models: Ridge, ElasticNet on the engineered feature set;

- a tree-based non-linear benchmark: XGBoost with early stopping.

Because the data are time-ordered, we use an expanding-window walk-forward split rather than random K-folds. Each of three folds reserves ≈15% of the timeline as a test window; training always uses all data up to the test start. We also used a 24-hour gap between train and test to reduce temporal dependency.

Models are evaluated with RMSE, MAE and R^2 on the full hourly distribution, and with additional RMSE/MAE metrics on the top 5% of hours by load (≈p95 tail). This reflects the dual objective: robust accuracy on typical hours and reasonable behaviour on high-load, tariff-relevant periods.

12. Key results and final model

Naive persistence provides a strong but purely reactive baseline: cross-fold results yield roughly $R^2 \approx 0.48$ and RMSE≈0.33 (in Gbps units). Seasonal naive performs worse, confirming that short-term inertia dominates simple daily repetition.

A regularised linear model with daily features, Ridge regression, substantially improves over Naive: mean test $R^2 \approx 0.74$ and RMSE≈0.23, with a very small train–test R^2 gap (~0.004). In relative terms, Ridge improves R^2 by about 27.6% and reduces RMSE by 21% vs the naive baseline. On the top 5% of hours (tail), Ridge achieves RMSE≈0.73 and MAE≈0.54, which is acceptable given the rarity and volatility of extreme peaks.

XGBoost, despite higher train R^2 (~0.83), underperforms Ridge on test ($R^2 \approx 0.71$) and shows a much larger generalisation gap, as well as unstable behaviour on peak hours. ElasticNet, after a grid search over α and l1_ratio, converges to a near-Ridge regime (l1_ratio≈0.1) and reproduces Ridge performance almost exactly, indicating that all 21 engineered features carry useful signal rather than obvious redundancy.

Our final model is Ridge regression trained on the weekly-extended feature set (daily + weekly lags). Adding weekly context yields a small but consistent gain: test $R^2 \approx 0.742$ (vs 0.740) and slightly lower RMSE (~0.2340 vs 0.2345). Coefficient inspection shows that daily drivers (`lag_1`, `sin_hour`, `cos_hour`, `lag_24`) remain dominant, while `lag_168` ("same hour last week") becomes a meaningful secondary signal. The improvement in metrics is modest, but it confirms that weekly structure is real and aligns this step more closely with the longer-horizon tariff and p95 goals.

13. Stakeholder-facing summary (short)

In practical terms, the model can flag the next hour as "likely heavier" or "likely lighter" than usual, based on recent traffic and typical daily/weekly patterns. It does not predict individual spikes with perfect precision, but it provides a reliable sense of where load is heading in the short term, and which hours are systematically riskier.

For operations and planning, this enables, for example:

•scheduling bandwidth-intensive jobs (backups, large syncs) away from hours that the model marks as high-risk;

•monitoring how often forecasts approach a chosen “comfort zone” relative to link capacity, as an early-warning signal for when tariff changes or upgrades may be needed;

•understanding which temporal regimes (time of day, day of week) and workload proxies tend to drive peaks.

The model is not an autopilot. It is a decision-support tool: engineers remain responsible for final capacity choices and for setting conservative buffers. In this capstone, the model demonstrates that next-hour peaks are predictable enough to be useful, while its limitations and uncertainty are transparent.

14. Technical limitations and future work

Technically, this project is a pilot on a single router with ≈5 months of hourly data. It does not capture annual seasonality, major routing changes, or rare crisis conditions. The feature set is intentionally compact (time features, lags, rolling statistics, host CPU, monthly VM growth proxies) and tailored to 1-step-ahead forecasting; it does not yet model longer horizons directly.

Future work includes:

•extending the approach to multiple routers / data centers, to test generalisability and allow network-wide capacity views;

•incorporating a longer history (12+ months) to capture yearly cycles and holiday effects;

•moving from 1-step-ahead to multi-step forecasting (e.g. 24–168 hours) and aggregating hourly predictions into monthly p95 estimates aligned with billing;

•systematically comparing Ridge to dedicated time-series and probabilistic methods (e.g. simple ARIMA/Prophet, quantile regression, or lightweight neural models) for the p95-focused use case;

•exploring additional operational features (e.g. routing policy flags, maintenance windows) if and when they become available.

Taken together, these steps would turn the current capstone from a robust short-horizon forecaster into a full capacity-planning pipeline that can directly support monthly tariff decisions under sustainability and reliability constraints.