

A Multi-Task Benchmark for Time Series Foundation Models on R&E Backbone Telemetry

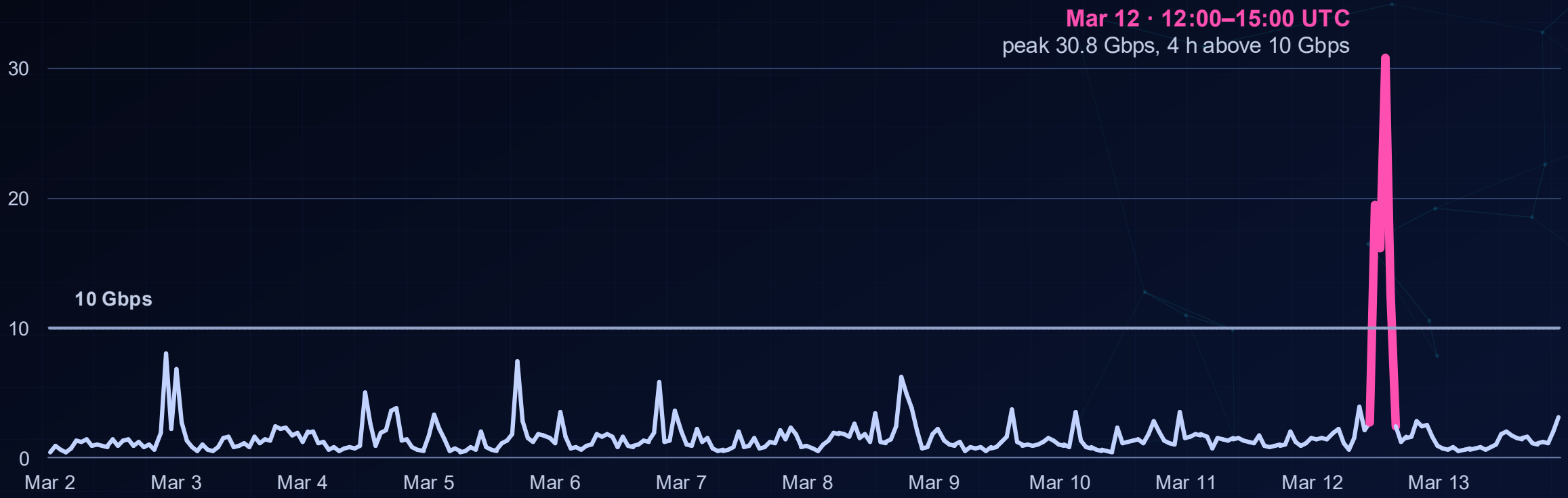
Gretel De la Peña, Hadi Sahin, Renata Frez, Italo Brito, Vasilka Chergarova, Luis Marin Vera, Jeronimo Bezerra, Julio Ibarra

Florida International University (AmLight) · RNP, Brazil

INDIS 2026 Workshop · SC26, Chicago

A real surge on a 100G backbone link

Outbound Gbps, hourly · one 100GE link of a production R&E backbone · March 2026



Could the NOC have known 24–48 hours earlier?

Plenty of data. Almost no labels.

Years

of SNMP telemetry archived by backbones

2–10

confirmed incidents per month on a backbone

≈ 0

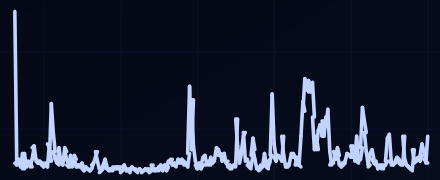
labels to train a supervised model

Supervised models need labels to learn

Time series foundation models (TSFM) are pre-trained on billions of series and forecast new data **without labels.**

THE ANSWER Freeze it, teach it

INPUT



Last 14 days
of SNMP counters

FOUNDATION MODEL

Large pre-trained network
learned from billions of time series

❄ FROZEN

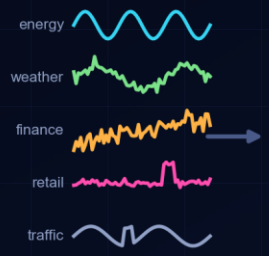
summary

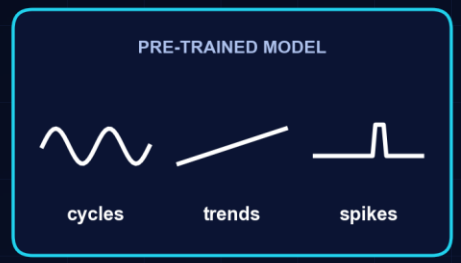
HEAD

Tiny head
one per question

Will it exceed 10 Gbps tomorrow?

yes / no

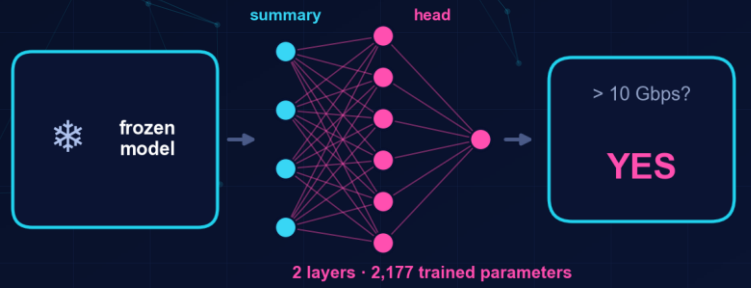




Foundation model
learned from billions of series



❄ **Frozen**
weights never change, it can't forget



Head
tiny add-on (2,177 params) for one question

0 human labels · < 1 s to train a head on a CPU · 8–120 ms per prediction, no GPU

THE DATASET

AmLight

R&E backbone

26 months
of hourly SNMP

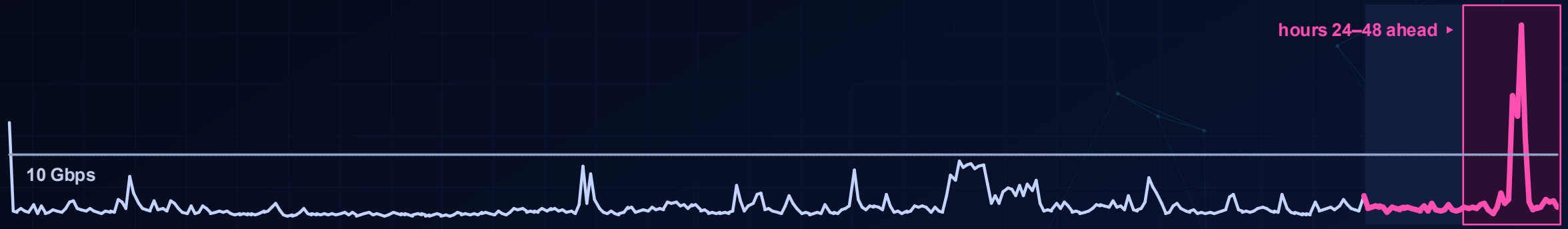
32 links
active 100GE

2 counters
bytes in / out

Public
anonymized

WINDOW = INPUT last 14 days of one link · 336 hourly values (Feb 25 – Mar 10)

NEXT 48 h after the window



A SCRIPT READS HOURS 24–48 AHEAD AND WRITES THE LABEL

Will it exceed 10 Gbps?

YES

peaked at 30.8 Gbps

No human effort: the archive already holds what happened next, so every past window can be labeled by a script.

EXPERIMENTS Benchmark

5 foundation models × **5** adaptation strategies × **4** NOC questions

THE 5 FOUNDATION MODELS WE COMPARE

TimesFM	Chronos	Moirai	TTM	PatchTST
Google	Amazon	Salesforce	IBM	no pre-trained.

Baseline: the best classical forecaster per task (ARIMA or Holt-Winters), always 24–48 h ahead.

Zero annotation: every label is computed from the 2 counters, so any network can rerun the benchmark.

EXPERIMENTS Five ways to adapt the model

SC26 | hpc unites. CHICAGO

CHANGE THE MODEL'S WEIGHTS

Full fine-tuning



100% trained

Retrain all weights on our telemetry

LoRA



≈ 0.6% trained

Add small trainable pieces to attention layers

ELoRA



≈ 1.2% trained

LoRA extended to more layers

KEEP THE MODEL FROZEN

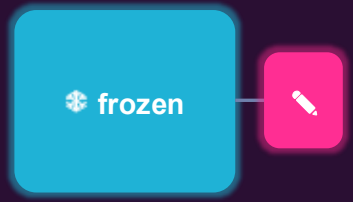
Zero-shot



0% trained

Use the model exactly as downloaded

ZS + MLP@n



only the head is trained · n = 5 … 500

Model stays frozen; only the small head is trained

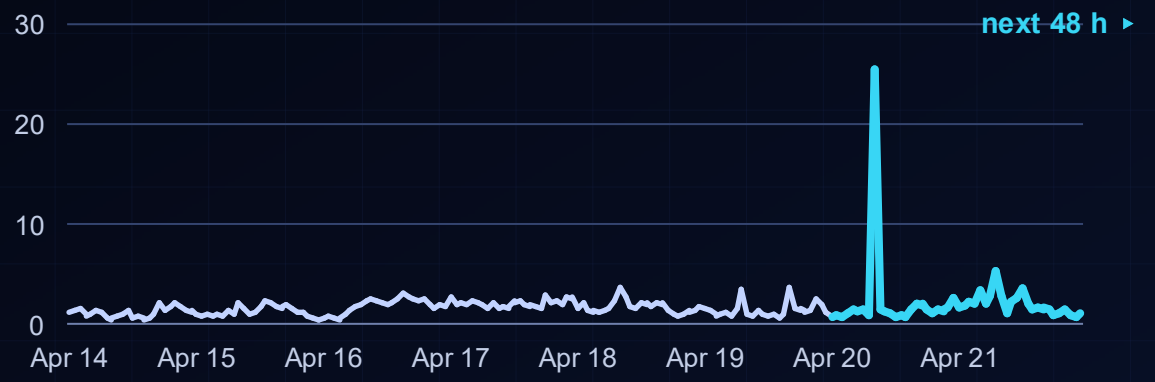
HOW TO READ THE PICTURES

- frozen model (weights unchanged)
- model weights being trained
- small head (trained)

Four questions a NOC asks, on real traffic

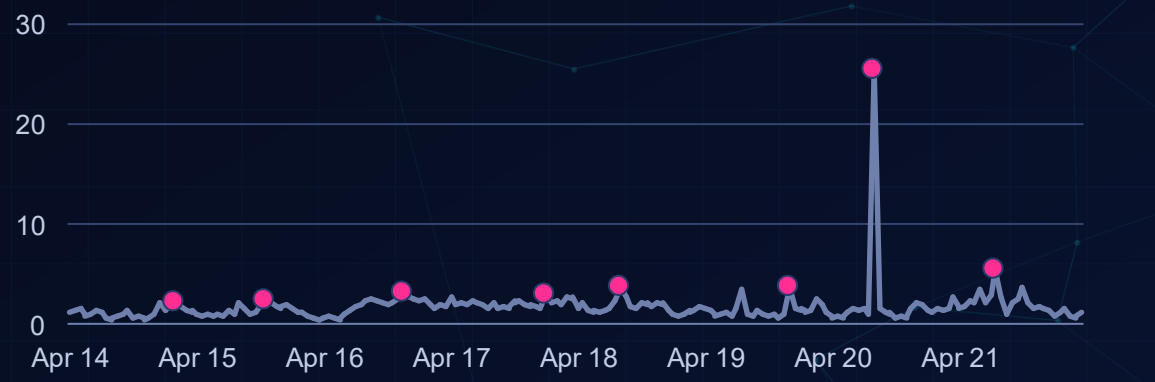
T1 How will the load evolve?

SHAPE & TIMING



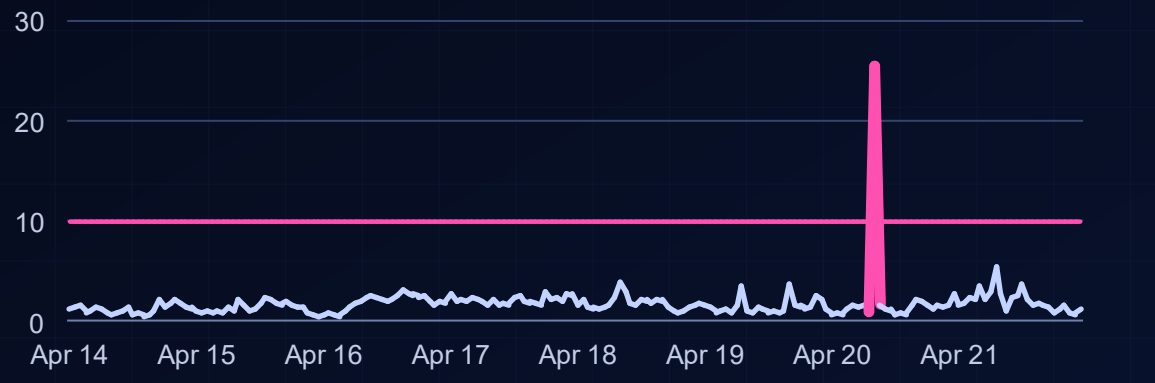
T3 When will the daily peak arrive?

SHAPE & TIMING



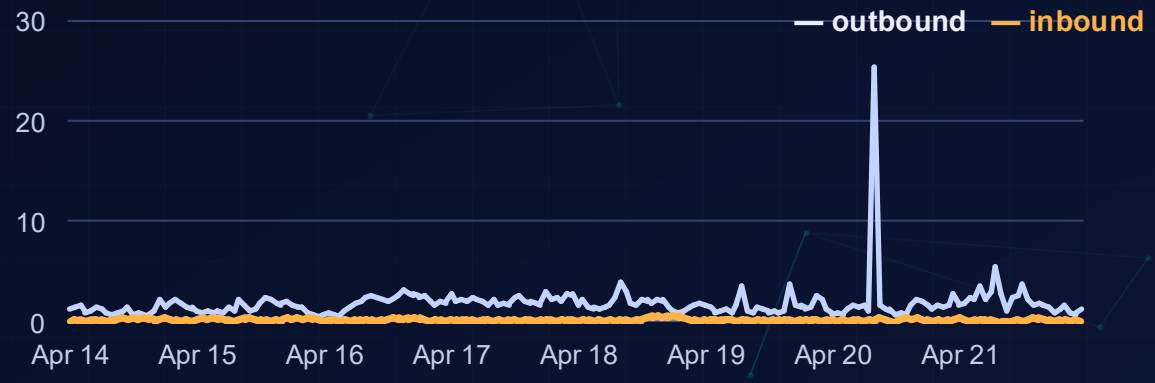
T2 Will it exceed 10 Gbps?

MAGNITUDE



T4 Mostly inbound or outbound?

MAGNITUDE



Same 100GE link, one week of hourly outbound Gbps, April 2026 (held-out test period). Forecasts are scored 24–48 h ahead.

LESSON 1 Zero-shot wins on shape & timing

SC26 | hpc unites. CHICAGO

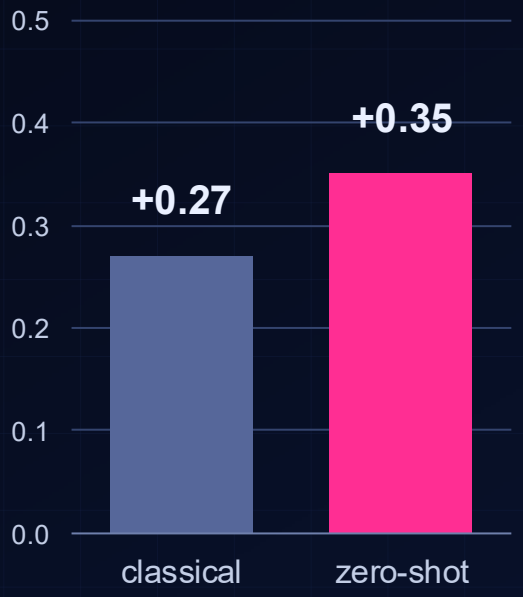
SHAPE & TIMING · foundation models win

T1 trajectory
skill ↑ higher is better



✓ zero-shot wins

T3 peak timing
skill ↑ higher is better



✓ zero-shot wins

Pre-trained models already know daily cycles, so timing comes for free.

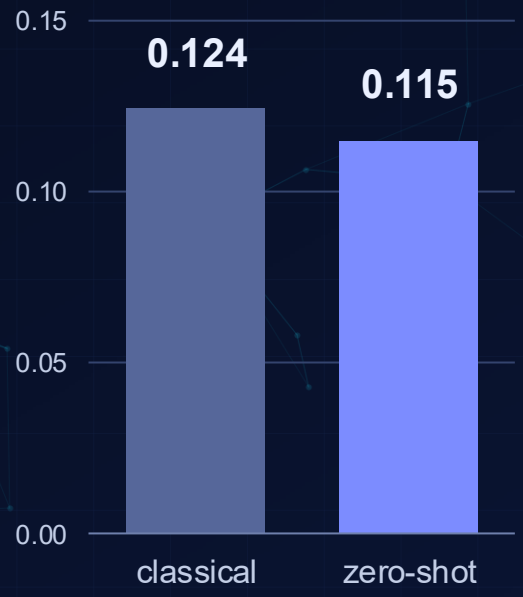
MAGNITUDE · classical holds on

T2 exceed 10 Gbps
AUROC ↑ higher is better



✗ zero-shot below

T4 in vs. out
error ↓ lower is better



= tie (not significant)

Load and direction barely change hour to hour (autocorrelation 0.92–0.95): repeating the recent past is hard to beat.

LESSON 2 Retraining can make them forget



Catastrophic forgetting

Forecasts flatten toward a constant: the daily pattern it knew is erased.

Not a tuning bug

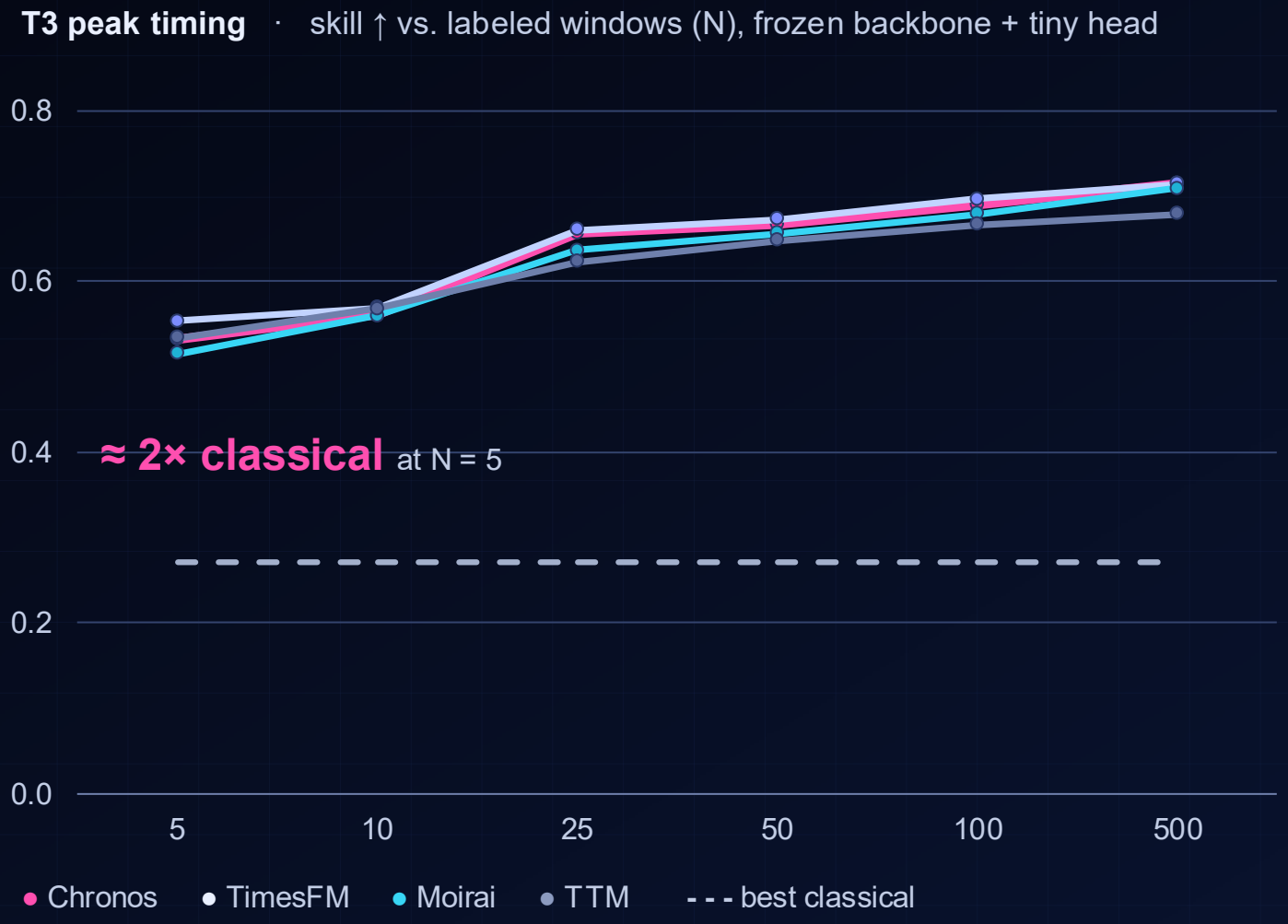
Clipped gradients, 10× lower learning rate: same collapse.

Architecture matters

Chronos' discrete tokens keep it robust under the same pipeline.

Across all 4 tasks, full fine-tuning never beat the frozen model.

LESSON 3 Five labeled windows are enough



5

windows for timing

Every model already doubles the best classical peak-timing skill: the head just reads out what the model knows.

500

windows for magnitude

≈ 3 weeks of one link: Chronos reaches 0.945 AUROC and beats ARIMA (0.940) on exceedance.

Which strategy for which question

What is your question about?

Shape & timing
how will it evolve? when?

T1 Load trajectory
Any model, zero-shot

no training

T3 Daily peak hour
Frozen model + tiny head

25 labeled windows

Magnitude
how much? which way?

T2 High-utilization event
Chronos + tiny head

500 labeled windows

T4 In vs. out
ARIMA is still enough

classical

RULE OF THUMB · NO NEED TO SWITCH MODELS TASK BY TASK

Very persistent signals
e.g. traffic direction → **a classical model is enough**

Everything else (trajectory, peak hour, high-utilization events)
one frozen model, preferably Chronos: its own forecast for trajectory, **+ a tiny head for other types of questions.**

In every case: never retrain all the weights.

The INTERNATIONAL CONFERENCE for HIGH PERFORMANCE COMPUTING NOV 15-20
CHICAGO, IL NETWORKING, STORAGE, & ANALYSIS



SC26
Chicago, IL | hpc unites.

SPONSORED BY IEEE COMPUTER SOCIETY TCHPC ACM Association for Computing Machinery sighpc

Don't retrain the foundation model. Freeze it and teach it your questions.

Code github.com/amlight/amlight-tsfm

Dataset doi.org/10.34703/gzx1-9v95/CAZJN8

Next replicate on other R&E networks: let's talk

Thank you! Questions?

Supported by NSF awards OAC-2029283, OAC-2029278, OAC-2537489