Network-Aware Caching in Distributed CMS XRootD Systems via Data-Driven Forecasting

Sarat Sasank Barla · Byrav Ramamurthy · Derek Weitzel

School of Computing, University of Nebraska–Lincoln, Lincoln, NE, USA

sbarla2@huskers.unl.edu · ramamurthy@unl.edu · dweitzel@unl.edu

INDIS 2026 - Innovating the Network for Data-Intensive Science

SC26 · Chicago, Illinois, USA · Sunday, November 15, 2026

Background

- The **Compact Muon Solenoid (CMS)** is a general-purpose detector at CERN's **Large Hadron Collider (LHC)**; its collisions generate petabyte-scale datasets.
- Data is distributed worldwide by the **Worldwide LHC Computing Grid (WLCG)** and served over **XRootD (XRD)**, imposing sustained load on **wide-area networks (WAN)**.
- **In-network caches (XCache)**, placed by operators such as the **Energy Sciences Network (ESnet)**, serve files locally and fetch from remote origins on a miss.
- The **AAA-ng** monitoring stream (*Any Data, Anytime, Anywhere*, next generation) records every **fallback read**: a wide-area transfer by construction.


Image: CMS Collaboration / CERN

Overview of the Work

- ***Problem:*** Deployed policies, **Least Recently Used (LRU)** and **Least Frequently Used (LFU)**, react only *after* a request arrives. They hold no model of future demand.
- ***Question:*** At what granularity is CMS access predictable at all?
- ***Measurement:*** Files are effectively non-reusable, **93.2%** of distinct files are requested exactly once per month. At a **6-hour** lag, top-10 hot-set overlap is **20.9%** for primary datasets, **7.2%** for directories, **0.25%** for files.
- ***Design:*** Forecast at the **primary-dataset** level, resolve down to dir/files, and couple the forecast to the cache so a wrong guess costs nothing.

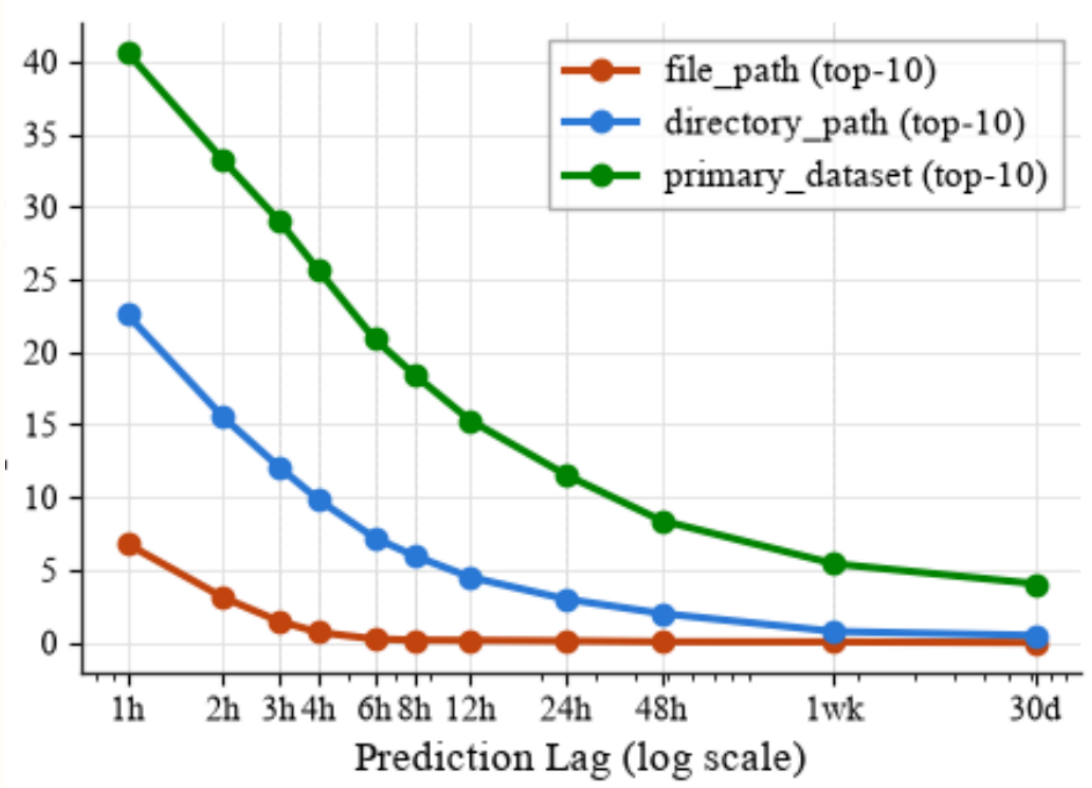


Fig. 1. Hot-set overlap vs. prediction lag at three path granularities (nine months; the ladder replicates at eight hosts across all four domains).

Methodology: A Two-Stage Predictive Pipeline



8,395,840
≈ 8.4 M session records

25 / 5 / 4
cache hosts / sites / domains

7,885,401
≈ 7.9 M CMS records kept

1.236 TB
LRU capacity in the replay

Granularity is measured, not assumed. Only primary datasets retain usable signal at operational prefetch horizons. The top-100 datasets cover ~88% of host request volume.

Prediction is scoped per serving host, not per client. The client decomposition discards signal rather than sharpening it.

Every record is a fallback read, a wide-area transfer, so bytes served locally subtract directly from backbone load.

Results

- **Forecast beats the reactive baseline.** The GRU overtakes persistence from **~12 hr**; at 24 hr, it reaches **31.2% / 33.3%** precision/recall vs **26.5% / 28.0%**. An inference-time blend lifts this to **35.7% / 39.8%** which is **+8.3 / +10.8 points** – at zero retraining cost.

- **The forecast has cache value.** Request-weighted hit rate on the predicted top-10 datasets reaches **79.2%** at 16 h against **72.2%** for the recency-equivalent policy: a **7.0 point** margin.

- **Coupling is what converts it into gain.** A side-cache beside an *unmodified* LRU raises the combined hit rate from **87.75%** to **89.38%** (**+1.63 pp**, 760 requests LRU alone would have missed) while forcing **zero** evictions.

- **Network impact: 115 GB** of wide-area traffic displaced at a single node over 23 days — **1.6%** of its WAN volume, at a sustained ~5 GB/day.

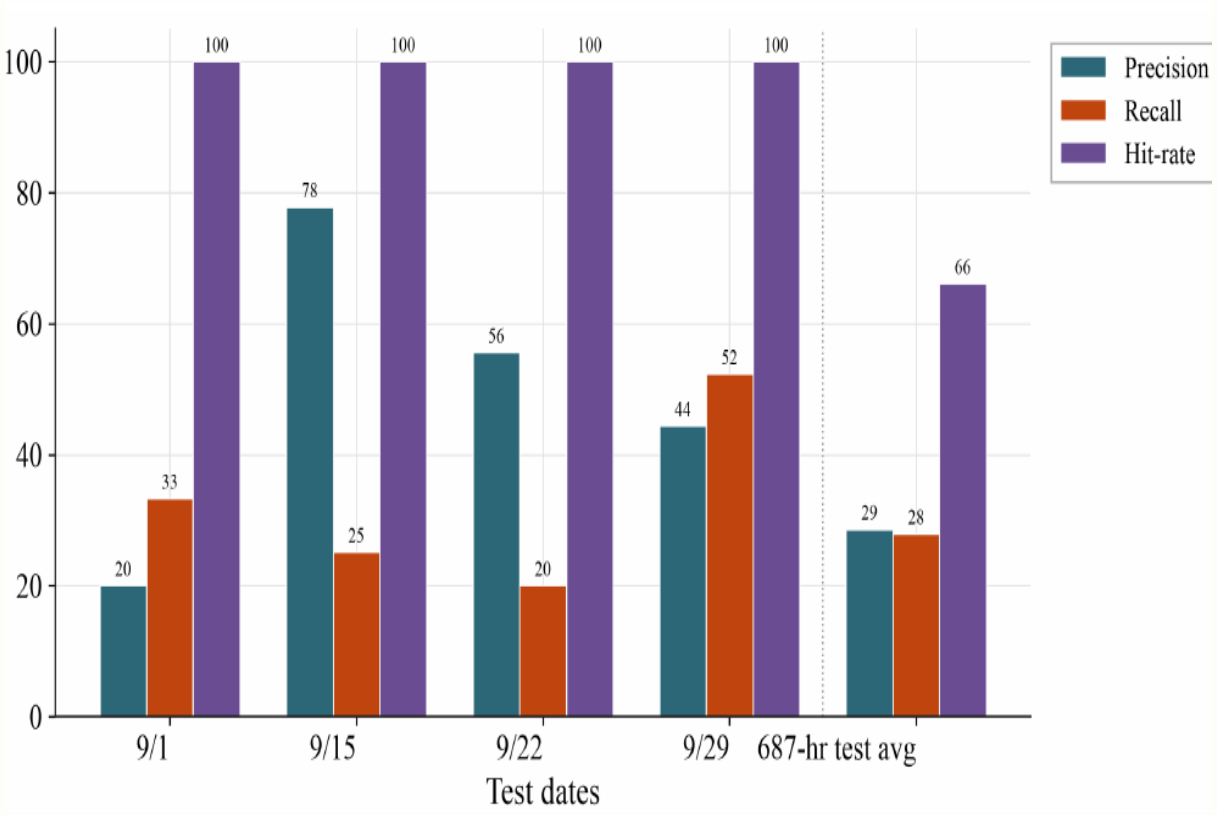


Fig. 3. Request-weighted hit rate of the predicted top-10 primary datasets versus the recency-equivalent policy across the held-out September test dates.

Conclusion

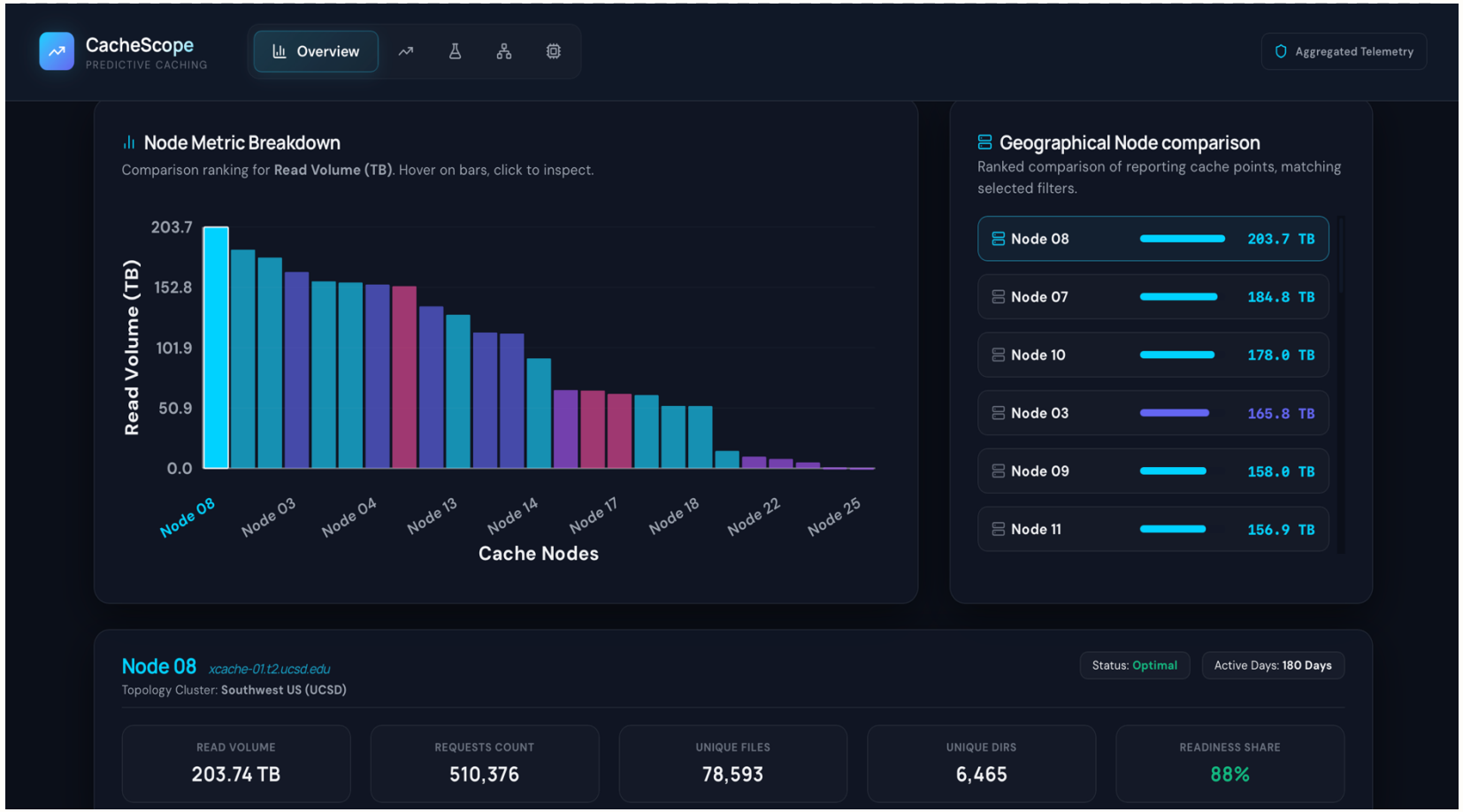
- ***Problem:*** In-network caches carrying CMS wide-area traffic are reactive, and predictive caching has never been validated at a granularity that demonstrably carries signal.
- ***Contribution:*** We establish the **primary dataset** as the operative forecasting granularity, build a **two-stage** GRU-plus-non-parametric pipeline, and couple it through a **side-cache** that cannot evict reactive content.
- ***Result:*** **+1.63 pp** combined hit rate over an **87.75%** LRU baseline, and **115 GB** (**1.6%**) of wide-area traffic displaced at one node in 23 days, measured on **8.4 M** production session records.
- ***Takeaway:*** The **coupling mechanism**, not forecast accuracy alone, determines whether prediction helps.

ACKNOWLEDGMENTS

Sponsored by the **U.S. Department of Energy (DOE)** under the **AI for High Energy Physics (HEP)** program

We thank **Diego Davila (UC San Diego)** for access to the U.S. CMS XRootD cache telemetry, and the U.S. CMS and ESnet operations teams.

UNL CacheScope – *Live Cache Telemetry Dashboard*



PROJECT

UNL CacheScope

Predictive caching telemetry for the U.S. CMS XRootD federation

VISIT

go.unl.edu/CMS-cache-info

TEAM

Sarat Sasank Barla

Byrav Ramamurthy

Derek Weitzel

School of Computing

University of Nebraska-Lincoln