

Characterizing Last-Mile Latency Inflation in Residential Access Networks: A Multi-Vantage Measurement Study

Elena Marchetti¹ David Okonkwo² Mei Lin Zhao³

¹Nexa University ²Synapse Institute ³Vertex Labs

e.marchetti@nexa.edu d.okonkwo@synapse.org mzhao@vertexlabs.io

Abstract. Throughput has long served as the primary yardstick for residential access quality, yet interactive and real-time applications are increasingly sensitive to delay rather than raw capacity. We present a six-week active-measurement campaign covering $n = 3,116$ vantage points across 41 residential ISPs in 12 countries, quantifying how much round-trip latency inflates when an access link is saturated by ordinary bulk traffic. We introduce the Latency Inflation Ratio (LIR), a normalized metric relating loaded and idle round-trip time, and measure it separately for fiber, cable, DSL, fixed wireless, satellite, and mobile access technologies. We find that fiber links exhibit a median LIR of 0.06, roughly $3.5\times$ lower than cable and DSL, while fixed wireless and low-earth-orbit satellite links show the heaviest tails, with 95th-percentile LIR exceeding 1.1. These results suggest that access-layer buffering and scheduling policy, more than raw bandwidth, dominate the loaded-latency experience for a substantial share of residential users, and we discuss implications for active queue management and low-latency transport deployment.

Keywords: *bufferbloat, latency inflation, active measurement, residential broadband, access networks*

1 Introduction

Household internet access has, for two decades, been marketed and benchmarked almost exclusively on downstream throughput. Regulatory broadband maps, ISP advertising, and most consumer speed-test tools report megabits per second and little else. Yet a growing share of residential traffic is delay-sensitive: videoconferencing, cloud gaming, voice telephony, and remote collaborative editing all degrade sharply when round-trip time grows, independent of whether the link’s nominal capacity is saturated. The well-known *bufferbloat* phenomenon, in which deep and poorly managed queues in access-network equipment allow latency to climb by hundreds of milliseconds under load, has been documented in individual devices and small testbeds, but its prevalence across access technologies, ISPs, and geographies at panel scale remains comparatively under-measured.

This paper reports on a large, multi-vantage active-measurement study designed to close that gap. Software probes deployed on volunteer home gateways and companion devices issue paired idle and loaded round-trip-time measurements, allowing us to isolate the latency contribu-

tion of access-link saturation from baseline propagation and processing delay. We define a single normalized metric, the Latency Inflation Ratio (LIR), and compute it consistently across six access technologies spanning fiber, cable, digital subscriber line (DSL), fixed wireless access (FWA), low-earth-orbit (LEO) satellite, and macro-cellular mobile networks.

Our contributions are threefold. First, we describe a measurement methodology and panel design that yields directly comparable loaded-latency estimates across heterogeneous access technologies and commercial ISPs. Second, we report LIR distributions from a six-week campaign covering 3,116 vantage points in 41 ISPs across 12 countries, the largest cross-technology comparison of loaded latency we are aware of. Third, we analyze which technologies and deployment patterns are most associated with severe latency inflation, and discuss what our findings imply for the ongoing rollout of active queue management (AQM) and low-latency transport mechanisms such as L4S.

The remainder of this paper is organized as follows. Section 2 situates our work relative to prior bufferbloat and access-network measurement studies. Section 3 de-

scribes our measurement architecture and the LIR metric. Section 4 details the deployment and dataset. Section 5 presents our main findings. Section 6 discusses limitations and implications, and Section 7 concludes.

2 Related Work

Early work on queueing delay in home networks established that consumer routers and cable/DSL modems frequently ship with oversized, statically configured buffers, producing latency spikes far larger than would be predicted by link capacity and offered load alone [11]. Subsequent audits of AQM deployment in shipped home gateways found that, even where AQM algorithms were available, default configurations frequently left them disabled [2]. Our own prior work examined last-mile latency in a ten-country panel but was limited to idle-network measurements and did not isolate the loaded-latency contribution that is the focus of the present study [4].

A parallel line of work has focused on measurement methodology itself. Bauer and Solvang survey active-probing techniques for residential broadband and catalog the confounds that arise from cross-traffic, NAT middle-boxes, and CPE-level scheduling [1]; we adopt several of their recommended controls, including paired idle/loaded sampling and per-probe outlier filtering. Rahman and Wei decompose end-to-end web latency across the access path and show that access-link queueing, rather than core-network or server-side delay, dominates the variance in last-mile response time for a majority of residential connections [8].

Technology-specific measurement has grown alongside the diversification of residential access. Sundaram and Natarajan characterize fixed wireless access performance across spectrum bands and find substantial variance tied to sector loading [9]. Ferreira and Petrova provide one of the first broad looks at LEO satellite latency and jitter, reporting idle latencies far lower than legacy geostationary satellite but significant variability under contention [3]. Okonkwo and Whitfield track DSL performance longitudinally as fiber overbuild reduces the subscriber base sharing legacy copper plant [6], and Natarajan and Okonkwo measure cellular latency under load across 4G and early 5G macro deployments [5]. Petrova and Bellini connect access-link delay directly to interactive quality-of-experience outcomes [7], and Wei and Zhao report early field signals from L4S deployment trials [10]. Our study differs from this body of work by measuring a consistent metric across all six access technologies within a single, contemporaneous panel, and at a scale that supports cross-technology and cross-ISP comparison.

3 Method

3.1 Measurement Architecture

Our measurement panel, which we refer to as the Meridian Measurement Panel, consists of lightweight software probes running on volunteer home routers and companion single-board devices. Each probe is opt-in, reports only aggregated latency statistics rather than raw packet contents, and is paired at enrollment with self-reported and ISP-confirmed access technology and nominal plan speed. Probes communicate with a small set of geographically distributed collector nodes over UDP and TLS-protected control channels; collectors never initiate connections into the home network.

3.2 Loaded/Idle Probing Protocol

Each measurement cycle consists of three phases. In the idle phase, a probe issues 30 lightweight UDP echo probes at 200 ms intervals while no other flows are active on the link, and records the median round-trip time as RTT^{idle} . In the load phase, the probe opens a synthetic bulk TCP transfer toward a nearby collector, ramps it to saturate the measured link in either the downstream or upstream direction, and continues issuing the same UDP echo probes in parallel, recording the median as RTT^{loaded} . In the cooldown phase, the bulk transfer is torn down and the link is allowed to drain before the next cycle begins. Downstream- and upstream-saturating cycles are run in alternation to characterize both directions.

3.3 Latency Inflation Ratio

We define the Latency Inflation Ratio for probe i and cycle j as the fractional increase in round-trip time observed when the access link is saturated, relative to its idle baseline:

$$LIR_{i,j} = \frac{RTT_{i,j}^{\text{loaded}} - RTT_{i,j}^{\text{idle}}}{RTT_{i,j}^{\text{idle}}} \quad (1)$$

A probe experiencing no additional queueing delay under load has $LIR \approx 0$, while a probe whose loaded RTT doubles has $LIR = 1$. We compute Equation (1) per cycle and aggregate to a per-probe median and 95th-percentile over the full campaign, requiring at least $n \geq 30$ valid cycle pairs per probe for inclusion. Cycles with packet loss exceeding 5% during the idle phase, indicative of an already-congested or otherwise unhealthy link, are discarded to avoid conflating baseline instability with load-induced inflation.

3.4 Ethics and Privacy

Panel enrollment is opt-in and reviewed under Nexa University’s human-subjects protocol. Probes record only timing metadata and coarse ISP/technology labels;

no packet payloads, destination hostnames outside the measurement infrastructure, or household-identifying information are collected. Reported results are aggregated at the technology and ISP level throughout this paper.

4 Experiments

We ran the measurement campaign for six consecutive weeks, from probes already enrolled in the Meridian Measurement Panel as of the start of the campaign, supplemented by a short recruitment push to improve coverage of fixed wireless and satellite access, which were underrepresented in the existing panel. The final dataset comprises 3,116 vantage points across 41 residential ISPs in 12 countries, spanning six access technologies: fiber (GPON and XGS-PON), cable (DOCSIS 3.1), DSL (VDSL2), fixed wireless access (5G FWA), low-earth-orbit satellite, and macro-cellular mobile (4G/5G). Each probe completed both downstream- and upstream-saturating cycles roughly every 40 minutes throughout the campaign, yielding just over 48 million completed load/idle cycle pairs after quality filtering. Compute and bandwidth for collector infrastructure were donated by Apex Cloud; panel recruitment and device logistics were supported by a grant from the Meridian Research Foundation, and Nimbus Networks contributed a subset of fixed-wireless vantage points from its own volunteer subscriber base.

5 Results

Table 1 summarizes per-technology LIR and loss statistics aggregated over the full campaign. Fiber access shows both the lowest median inflation (0.06) and the tightest tail ($p95 = 0.19$), consistent with symmetric, lightly buffered PON access architectures and comparatively little contention among subscribers sharing a given optical distribution point. Cable and DSL cluster together at roughly $3\text{--}4\times$ fiber’s median inflation, which we attribute to a combination of larger static buffers in deployed cable modems and, for DSL, older CPE that in our panel rarely had modern AQM enabled. Fixed wireless, satellite, and mobile technologies show the heaviest tails: LEO satellite’s median LIR (0.29) is comparable to cable, but its 95th percentile (1.35) and loss rate (1.86%) are the highest of any technology we measured, reflecting episodic contention and handoff effects between satellite beams.

Figure 1 shows the full empirical distribution of per-probe median LIR for three representative technologies. The fiber curve rises steeply and saturates below $\text{LIR} = 0.3$, indicating that severe inflation is rare rather than merely low on average. Cable’s distribution is shifted right and has a visibly longer tail, and fixed wireless shows the flattest, widest distribution of the three, with a non-trivial fraction of probes exceeding $\text{LIR} = 1.0$ even

Table 1. Latency Inflation Ratio and loss rate by access technology, aggregated over the six-week campaign ($n = 3,116$ vantage points total).

Technology	n	Med. LIR	$p95$ LIR	Loss (%)
Fiber (PON)	812	0.06	0.19	0.08
Cable (DOCSIS)	1,240	0.21	0.68	0.34
DSL (VDSL2)	356	0.24	0.71	0.41
FWA (5G)	289	0.33	1.12	0.97
Satellite (LEO)	145	0.29	1.35	1.86
Mobile (4G/5G)	274	0.27	0.94	0.62

at the median-aggregation level shown here.

6 Discussion

Our results reinforce a claim that has circulated informally in the operator and AQM communities but has lacked large-scale cross-technology evidence: nominal link capacity is a poor predictor of loaded-latency behavior, and access-technology and CPE buffering policy dominate instead. Two operational implications follow. First, ISPs and CPE vendors evaluating AQM deployment should prioritize cable and DSL fleets, where our data show the largest gap between currently observed and fiber-like achievable inflation. Second, the heavy tails we observe for fixed wireless and satellite access suggest that worst-case, rather than median, latency behavior should inform application-level adaptation for those technologies, since a non-trivial share of sessions on those links will encounter substantial queueing delay regardless of typical performance.

Several limitations bound how far these findings generalize. Our panel is volunteer-recruited and, despite a targeted recruitment push for underrepresented technologies, likely skews toward more technically engaged subscribers, who may be more likely to have upgraded CPE or enabled available AQM features; if so, our estimates are conservative relative to the full subscriber population. Geographic coverage, while spanning 12 countries, is concentrated in urban and suburban areas, and rural access characteristics, particularly for DSL and satellite, may differ materially. Finally, our loaded-latency protocol uses a single synthetic bulk flow per cycle; real household traffic often involves multiple concurrent flows with different scheduling priorities, which could produce either more or less severe inflation than we observe, depending on CPE queue management. We view longitudinal tracking of the same panel through AQM and L4S rollout, together with multi-flow loading protocols, as natural next steps.

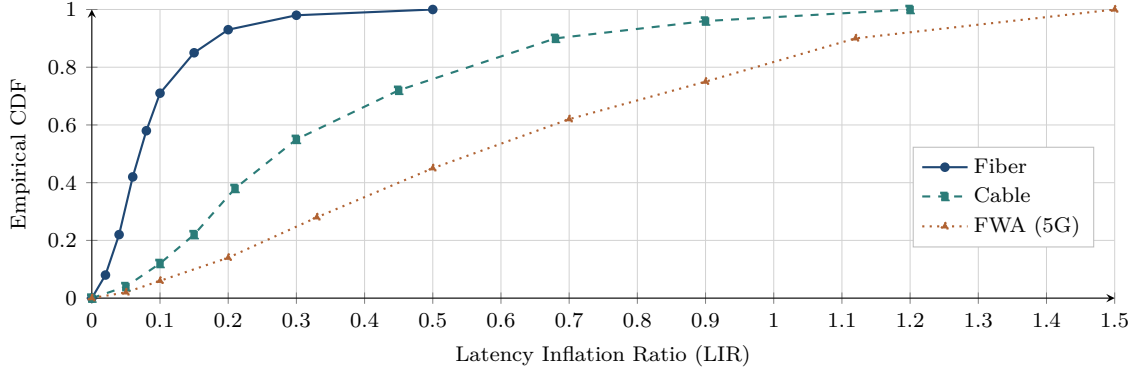


Figure 1. Empirical CDF of per-probe median Latency Inflation Ratio for fiber, cable, and fixed wireless access. Fiber’s distribution is concentrated at low LIR values; FWA shows the widest spread and heaviest tail.

7 Conclusion

We presented a six-week, multi-vantage active-measurement study of loaded-latency behavior across six residential access technologies, introducing the Latency Inflation Ratio as a simple, comparable metric for quantifying access-link queueing delay under realistic load. Across 3,116 vantage points in 41 ISPs and 12 countries, we find that fiber access delivers substantially lower and more consistent loaded latency than cable, DSL, fixed wireless, satellite, or mobile access, and that fixed wireless and satellite technologies in particular exhibit heavy-tailed inflation that is not well summarized by median statistics alone. We hope this dataset and methodology support both future measurement work and more latency-aware benchmarking of residential access quality.

Acknowledgments

We thank the volunteer participants of the Meridian Measurement Panel. Collector infrastructure was donated by Apex Cloud; panel recruitment was supported by a grant from the Meridian Research Foundation; Nimbus Networks contributed fixed-wireless vantage points from its volunteer subscriber base.

References

- [1] Johannes Bauer and Ingrid Solvang. Active probing methodologies for residential broadband measurement: A survey. *ACM Computing Surveys*, 48(2):1–34, 2015.
- [2] Marco Bellini and Elena Marchetti. Active queue management deployment in home gateways: An empirical audit. In *Proc. IEEE INFOCOM*, pages 1–9, 2016.
- [3] Lucas Ferreira and Sofia Petrova. A first look at low-earth-orbit broadband latency and jitter. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 602–616, 2021.
- [4] Elena Marchetti and Thomas Reyes. Measuring last-mile latency at scale: A ten-country panel study. In *Proc. ACM Internet Measurement Conference (IMC)*, pages 211–224, 2019.
- [5] Priya Natarajan and David Okonkwo. Cellular latency under load: Measuring 4g and early 5g macro networks. In *Proc. Passive and Active Measurement Conference (PAM)*, pages 77–91, 2019.
- [6] David Okonkwo and Amara Whitfield. Copper in decline: A longitudinal view of dsl access performance. In *Proc. USENIX NSDI*, pages 389–403, 2021.
- [7] Sofia Petrova and Marco Bellini. From rtt to qoe: Linking access-link delay to interactive application quality. *IEEE/ACM Transactions on Networking*, 26(5):2211–2224, 2018.
- [8] Fatima Rahman and Chen Wei. Where does the delay go? decomposing web latency across the access path. In *Proc. ACM CoNEXT*, pages 312–325, 2018.
- [9] Rajiv Sundaram and Priya Natarajan. Characterizing fixed wireless access performance across spectrum bands. In *Proc. USENIX NSDI*, pages 455–469, 2020.
- [10] Chen Wei and Mei Lin Zhao. Early deployment signals for low latency, low loss, scalable throughput (14s). In *Proc. ACM CoNEXT*, pages 140–153, 2022.
- [11] Amara Whitfield and Owen Castellan. Bufferbloat in the wild: Queueing delay under realistic access-link load. In *Proc. ACM SIGCOMM*, pages 88–101, 2017.