allot
See. Control. Secure.

Allot Network Intelligence

Why AI Traffic Demands a New Telco Network Strategy

July 2026

Table of Contents

Why AI Traffic Demands a New Telco Network Strategy

1. Intro and key takeaway

AI, including LLMs and generative AI, is not just increasing Internet traffic volumes. It is changing the very shape of traffic itself: where it flows, how it is generated, when it bursts, and which network resources it stresses. For network operators, these shifts affect capacity planning, backbone investment, peering strategy, caching architecture, and Quality of Experience (QoE) management.

Industry analyses increasingly point to the same conclusion: global network traffic is changing in character, not only in volume, as AI-driven applications spread. Video streaming remains the largest contributor to raw bandwidth, but AI introduces traffic classes with different directionality, burst patterns, cacheability, and latency sensitivity. For telcos, this makes AI traffic a planning issue, not only an application trend. Operators need to reassess the role of DNI, traffic shaping, packet prioritization, and AI-aware analytics in multi-year network evolution.

Executive takeaway: The core message for telcos is clear: AI traffic should not be treated as just another source of bandwidth growth. It changes traffic directionality, concentration, cacheability, latency sensitivity, and the relationship between network KPIs and user experience. Operators need AI-aware traffic visibility, policy control, and analytics to manage this shift without overbuilding blindly or degrading human-facing services.

This requires AI-aware traffic intelligence, classification, policy control, and analytics capabilities that can distinguish between human-facing AI usage, background machine traffic, AI crawlers, and conventional application traffic.

2. User-Driven Volume Increase

New technologies tend to increase traffic volumes, and AI could accelerate this effect in specific parts of the network. A conventional search session may lead a user to open only a few pages, while AI systems can decompose a query into multiple sub-queries, retrieve many candidate sources, and synthesize a response from distributed compute and storage resources. This can generate significantly more back-end traffic per user intent, even when the user sees only one answer.

While this increase in traffic volume is ultimately user-driven, much of the incremental load may not scale directly with last-mile user demand. Instead, a significant share can appear as east-west traffic between data centers, as AI systems decompose queries, execute parallel retrieval, rank candidate information, and synthesize responses across distributed compute and storage resources.

This architectural shift has important implications for network planning. First, traffic growth becomes less tightly coupled to access bandwidth growth. Even when access-network traffic grows at a moderate pace, data-center and inter-data-center traffic can grow faster because of AI inference, retrieval-augmented generation, and model orchestration workloads. Second, AI-driven east-west traffic has characteristics that differ from conventional user traffic. It can be highly bursty and synchronized across compute nodes. These flows are not easily cacheable and often offer limited locality, reducing the effectiveness of traditional CDN and caching strategies. The result can be additional latency and congestion, especially on data-center-interconnect (DCI) links, with possible downstream effects on adjacent network domains.

Finally, operators that focus primarily on last-mile utilization or aggregate downstream volume risk underestimating the scale and urgency of investments required in metro, core, and DCI infrastructure. As AI adoption accelerates, the dominant constraint on service quality may no longer be access bandwidth but rather the capacity, latency, and resilience of the networks supporting high-rate east-west traffic.

3. Concentration and the Collapse of Click-Through Traffic

3.1. Traffic Fan-out reduction

A central trend driven by generative AI is the concentration of answers. End users increasingly receive synthesized responses directly from AI systems such as ChatGPT, Microsoft Copilot, or Google Gemini rather than clicking through to multiple websites. BAIN & COMPANY REPORTS that 80% of consumers rely on zero-click results at least 40% of the time. From a network perspective, the important shift is that the visible user-side fan-out may shrink, while more activity is concentrated behind AI platforms and their supporting cloud infrastructure.

Historically, a single user query resulted in:

- DNS lookups for many domains
- Multiple TCP or QUIC connections to diverse content servers
- Further connections related to ads, images, tracking and scripts

With AI-generated answers, the user-facing traffic pattern becomes more concentrated:

- Destination servers belong to a small number of AI companies (OpenAI, Google, Anthropic)
- Long-lived, encrypted connections to a few large autonomous systems.

This reduces destination diversity and steers telco traffic toward large, centralized platforms and CDNs, increasing backbone and interconnect concentration.

3.2. Impact on Websites

This architectural change has measurable effects on web publishers. SEVERAL STUDIES INDICATE that a large share of searches now end without a click-through, especially when AI summaries satisfy the user's intent directly. Bain & Company reported that, as of February 2025, 60-70% of searches resulted in no click-through, and other industry analyses point to similar ranges when AI Overviews are shown.

For telecom operators, however, the real concern isn't the drop in publisher traffic—it is how this behavior alters the network itself. Instead of users browsing a diverse web, traffic is consolidating. This manifests as:

a. **Fewer distinct HTTP transactions** as users stop clicking through to individual websites.
b. **A sharp drop in cacheable objects** because static web pages are bypassed.
c. **Heavy reliance on a small number of AI service endpoints**, shifting network demand toward massive, centralized data centers.

Because AI-search behavior is evolving quickly, these figures should be read as directional rather than definitive. The exact percentage varies by study, query type, geography, and measurement method, but the broader trend is consistent: AI summaries and zero-click experiences reduce the number of visible referrals to many websites.

4. Human User Traffic vs. AI Inference and MCP Traffic

As Allot has consistently observed, video streaming accounts for the largest share of downstream bandwidth in access networks (typically 60-70%, depending on geography). Such an Internet strongly benefits from:

- CDN edge caching
- Downlink-heavy flows
- Relatively predictable connection duration times.

AI inference traffic behaves quite differently. Although raw bandwidth per request is modest, AI applications typically generate:

- Short user queries (text, images, or audio)
- Compute-bound responses, often streamed token-by-token
- Long-lived TCP or QUIC sessions with variable completion times.

A GOOGLE CLOUD BLOG notes that generative AI inference differs sharply from traditional web applications

- Traditional web applications have predictable request/response times measured in milliseconds, while AI inference exhibits highly variable request/response times, often seconds to minutes
- A single LLM query may consume 100% of a GPU/TPU, unlike traditional web request processing, which consumes a fraction of a server CPU
- Requests are not parallelizable and often wait for available compute, not network capacity
- Caching is largely ineffective because responses are unique per query

An emerging traffic category, alongside that of AI chatbots, is the embedding of AI capabilities into legacy web applications and services. The traffic from such AI-augmented applications falls between conventional web traffic and the previously discussed traffic. Upon introduction of AI a user may begin experiencing increased latency and may blame the network for apparent sluggishness, although new capabilities may offset the QoE impact.

The main takeaway is that while video streaming traffic maps well to hierarchical caching, AI user traffic doesn't, reducing the effectiveness of both CDN edge caching and operator-deployed transparent caches.

5. Massive Traffic Volumes Generated by AI Crawlers

An often-underappreciated traffic source is AI data ingestion. AI models depend on continuous large-scale crawling of the public web for:

- Initial training
- Periodic re-training
- Retrieval-augmented generation (RAG) corpus refresh

AI crawlers, such as GPTBot, Meta-ExternalAgent, and ClaudeBot, are pure machine-to-machine automated systems, fetching vast numbers of pages with no human in the loop.

AI crawlers now account for a rapidly growing share of bot-initiated traffic. Cloudflare Radar data shows that:

- AI crawler traffic often exceeds that of traditional search engine crawlers during peak periods
- From July 2024 to July 2025, GPTBot traffic increased by 147%
- In the same period, Meta-ExternalAgent traffic increased by 843%
- AI bots crawl at scale without returning equivalent referral traffic, unlike traditional search crawlers.

CLOUDFLARE EXPLICITLY FRAMES THIS as a new class of Internet traffic.

AI crawler workloads generate:

- Large-scale machine-to-machine traffic

- Strong temporal bursts
- Concentration on specific content-heavy domains

For telcos, this increases uplink traffic, strains peering links, and complicates bot mitigation and traffic classification.

When AI crawler behavior becomes abusive, anomalous, or destabilizing, Allot NetProtect can help operators detect and mitigate abnormal traffic patterns, including DDoS, L7 attacks, and botnet-driven activity. This framing is important: AI crawlers are not automatic attacks, but excessive automated behavior can create operational impact that requires detection, control, and mitigation.

6. AI and User QoE

Traditional QoE degradation is often dominated by:

- Last-mile congestion
- Packet loss
- RTT inflation

AI inference QoE is dominated by a completely different bottleneck - AI response latency that is almost completely governed by computational delay rather than network delay. While inference workloads are individually lighter than training jobs, the explosive growth in overall inference volumes places sustained pressure on GPU and network availability, leading to higher user-perceived latency.

From a telco perspective:

- Network KPIs may appear healthy
- Yet users still experience disappointingly slow responses

From a telco perspective, this creates a measurement gap: network KPIs may appear healthy, users may still experience slow AI responses, and traditional QoE analytics may no longer explain the full user experience. This weakens the correlation between network performance and perceived service quality.

7. AI Agents: A Future Traffic Pattern Shift

AI agents, autonomous systems that can act on a user's behalf across applications and services, are still an emerging trend rather than a major source of traffic today.

The behavior of an AI agent differs sharply from that of human users:

- Interactions are API-heavy rather than UI-driven
- Traffic volume does not generally obey diurnal constraints
- Large fractions of the traffic are machine-to-machine (east-west traffic)
- Activity is triggered by machine events, not by human reading speeds
- Activity bursts are driven by software heartbeat times and AI compute availability
- Agent-driven traffic could introduce synchronized, cross-application bursts

Furthermore, agents may communicate with other agents, creating traffic flows even more radically different from human-generated ones.

For telcos, this means:

- Less predictable traffic cycles
- Increased east-west and inter-DC flows
- New challenges in anomaly detection and capacity planning.

8. How Allot Can Help Telcos Manage AI Traffic

8.1. The Crucial Role of DNI-driven Policy

As we have seen, AI traffic is both behaviorally distinct from human-driven traffic and subject to requirements that differ from those that define acceptable Quality of Experience (QoE) for human users. At the same time, AI-generated and AI-consuming traffic does not exist in isolation: it traverses the same access networks, aggregation routers, peering links, CDNs, data-centers, and middleboxes as conventional human sessions. This fundamental coexistence makes undifferentiated treatment increasingly inefficient and, in some cases, actively harmful to both network efficiency and end-user experience.

DNI-driven policy, therefore, becomes a vital capability. Without fine-grained traffic classification, networks are effectively blind to the impact of AI traffic. Treating AI and human traffic equally is no longer viable when long-lived, high-throughput machine-generated flows compete with latency-sensitive human interactions, or when aggressive AI crawlers consume disproportionate amounts of network resources.

At the foundation of this approach is Allot DART2.0, Allot's application identification and classification engine. DART2.0 recognizes dozens of AI-related applications, and its coverage is updated regularly as new services emerge. Within DART2.0, Allot TraffiCognition provides AI-driven classification technology that helps classify ECH traffic, where conventional visibility approaches are no longer viable. These capabilities enable telcos to distinguish between materially different traffic categories, including:

- Human sessions, such as web browsing, messaging, and popular apps
- Interactive AI inference sessions, where a human is directly engaged with an AI system
- Services that utilize AI, but present themselves as conventional apps
- AI inference, AI search, and Model Context Protocol (MCP) traffic, where applications connect models to external tools, data sources, and services
- AI crawlers, typically characterized by high request rates, minimal think time, and machine-to-machine communication patterns

Note that interactive inference sessions may exhibit traffic characteristics resembling machine behavior, such as bursty token streams or atypical flow durations, while still requiring human-grade QoE in responsiveness and reliability. Conversely, background AI agents or crawlers may tolerate higher latency or throttled throughput without any user-visible degradation.

DNI rule sets must be continuously updated to reflect not only entirely new AI services but also new AI-driven behaviors within well-known protocols and platforms.

By providing accurate, real-time visibility into AI traffic categories, Allot DART2.0 facilitates the application of appropriate differentiated policies. These may include prioritizing latency-sensitive interactive AI sessions, deprioritizing bulk or background AI activity during congestion, or detecting and mitigating abusive AI traffic before it impacts network stability.

Beyond traffic management, the resulting visibility and analytics support longer-term strategic functions. Trend analysis of AI traffic growth and composition informs capacity planning and interconnect optimization; the correlation between application types and network load assists in designing novel service plans; and, along with geographic and regulatory awareness, it contributes to enforcing data sovereignty and compliance principles. In this sense, DNI-driven policy is not only a defensive mechanism but a key enabler for sustainable network evolution in the era of AI.

8.2. Traffic Management in the Era of AI

Traffic shaping is not a magic bullet. It cannot reduce LLM computational delay, eliminate queuing inside hyperscaler data centers, or make fundamentally uncacheable AI interactions cache-friendly. Nevertheless, when applied appropriately, traffic shaping remains a powerful tool for enabling stable coexistence between AI traffic and legacy, human-driven sessions. By mitigating problematic AI traffic patterns, shaping contributes to fairness, predictability, and fault containment.

This is where Allot SmartTraffic fits naturally. SmartTraffic translates classification into policy enforcement: prioritizing latency-sensitive interactive AI sessions, rate-limiting excessive crawler activity, deprioritizing non-interactive background AI traffic during congestion, and protecting human-facing applications from AI-driven bursts.

Unchecked crawlers can generate sustained request rates that stress uplinks, peering links, or shared CDN ingress points, especially when many crawlers target the same content-heavy domains at the same time.

Properly applied, traffic shaping can support coexistence with AI by:

- Containing M2M bursts that are poorly handled by legacy congestion models. Many AI workloads generate synchronized request spikes (for example, periodic retrieval or model refresh behavior) that do not react gracefully to packet loss or ECN signals, amplifying short-lived congestion into wider instability.
- Protecting access networks from asymmetric uplink saturation. AI traffic increasingly exhibits non-traditional directionality behavior, including heavy upstream components driven by complex prompts, lengthy contexts, and telemetry. This is particularly problematic in consumer and SMB access networks engineered for downstream-heavy human use.
- Smoothing burstiness to avoid micro-congestion cascades. Even when average bandwidth consumption is moderate, highly bursty AI traffic can cause transient queue buildup, buffer oscillations, and head-of-line blocking that degrade latency for unrelated, latency-sensitive flows.

From an operational perspective, the objective of traffic shaping is not to penalize AI traffic, but rather to limit its impact during resource contention, ensuring that transient or background AI activity does not crowd out interactive human sessions.

Particular caution is warranted when dealing with interactive AI inference sessions. Although these sessions may exhibit machine-like characteristics, such as rapid token streaming or atypical flow durations, their QoE expectations are like those of human traffic. Excessive shaping in this category risks introducing latency inflation or response jitter that is directly perceptible to end users, undermining the very services shaping is meant to protect.

This reinforces the importance of DPI. Shaping that is blind to context quickly becomes counterproductive; shaping that is informed by DPI-driven categorization can remain both effective and proportionate. In practice, this means identifying and controlling AI traffic that is clearly non-interactive, in the background, or abusive, while prioritizing human-facing flows.

In summary, DPI-driven traffic shaping is an essential tool for preserving network stability, protecting human QoE, and preventing pathological AI traffic patterns from triggering disproportionate downstream effects. In the AI era, traffic management is no longer a simple matter of “gold” subscribers versus “bronze” subscribers, or of real-time connections versus browsing. Instead, policy must ensure that human interactive traffic (including AI-assisted human-facing interactions) is not negatively impacted by AI crawlers, batch inference, or other M2M AI traffic, while still allocating sufficient resources to AI workloads to avoid outright starvation.

8.3. Analytics in the Era of AI

In the AI era, analytics must go beyond mere visibility. Meaningful analytics must translate raw observations into QoS-relevant Key Quality Indicators (KQIs) that reflect how different traffic classes actually behave and impact the network. Without KQIs that capture latency sensitivity, burst tolerance, asymmetry, and congestion contribution across human, human-facing AI, and M2M AI traffic, operators are left with dashboards that describe traffic but do not support decisions. Effective analytics, therefore, bridges visibility and action, quantifying service quality in a way that directly informs policy, prioritization, shaping, and long-term network planning.

Unfortunately, traditional QoS indicators, such as throughput, packet loss, and round-trip time, are no longer sufficient to describe the service quality experienced by the new AI traffic classes. AI introduces new dynamic flow sensitivities, e.g., responsiveness and jitter of interactive exchanges, burstiness, upstream vs. downstream asymmetry, and time to create short-lived M2M flows. Flows encountering network conditions that appear acceptable under legacy KQIs may nevertheless experience severe QoE degradation.

Operators need to define AI-relevant KQIs, and analytics providers need to measure these metrics and track their influence on different AI traffic categories. The goal is to build an analytics layer that connects DPI-level traffic understanding with relevant metrics, so that prioritization, shaping, and capacity decisions may be driven by the KQIs that actually matter for each traffic class.

Allot Visibility addresses this need by helping operators move from raw traffic visibility to actionable AI traffic intelligence. It can support tracking AI traffic growth trends, identifying where AI traffic concentrates, measuring the impact of bursts and asymmetry, and informing capacity planning, peering strategy, and policy decisions.

9. Summary

AI is reshaping Internet traffic in respects that go far beyond mere volume growth:

- **Answer concentration** collapses user fan-out into a small number of AI platforms, reducing destination diversity and reshaping traffic topology.
- **AI inference traffic** differs fundamentally from human video-centric traffic, being query-driven, compute-bound, and largely uncacheable.
- **AI crawlers** generate large, bursty, upstream-heavy machine traffic that stresses peering and complicates traffic management.
- **User QoE** is increasingly dominated by computational latency rather than network latency, weakening the relevance of traditional QoS indicators.
- **AI agents**, though nascent today, point toward a future of machine-driven traffic patterns completely decoupled from human behavior.

AI does not just add new application categories; it fundamentally reshapes how traffic is generated, concentrated, consumed, and experienced across the Internet.
For telcos, these trends require a re-evaluation of network planning assumptions, caching strategies, backbone capacity, peering design, and analytics collection. As AI traffic grows, operators that treat all encrypted cloud traffic as indistinguishable will have less visibility and fewer policy options. DPI-driven, AI-aware traffic management is becoming a core operational capability rather than an optional value-added service.

Allot helps operators move from undifferentiated encrypted traffic to AI-aware traffic intelligence, policy control, and analytics. DART and TraffiCognition provide the classification foundation, SmartTraffic turns classification into traffic management and policy enforcement, Visibility supports AI traffic trends, KQIs, and planning, and NetProtect helps detect and mitigate DDoS, L7, botnet, and anomalous traffic behavior where automated AI-related activity becomes harmful.