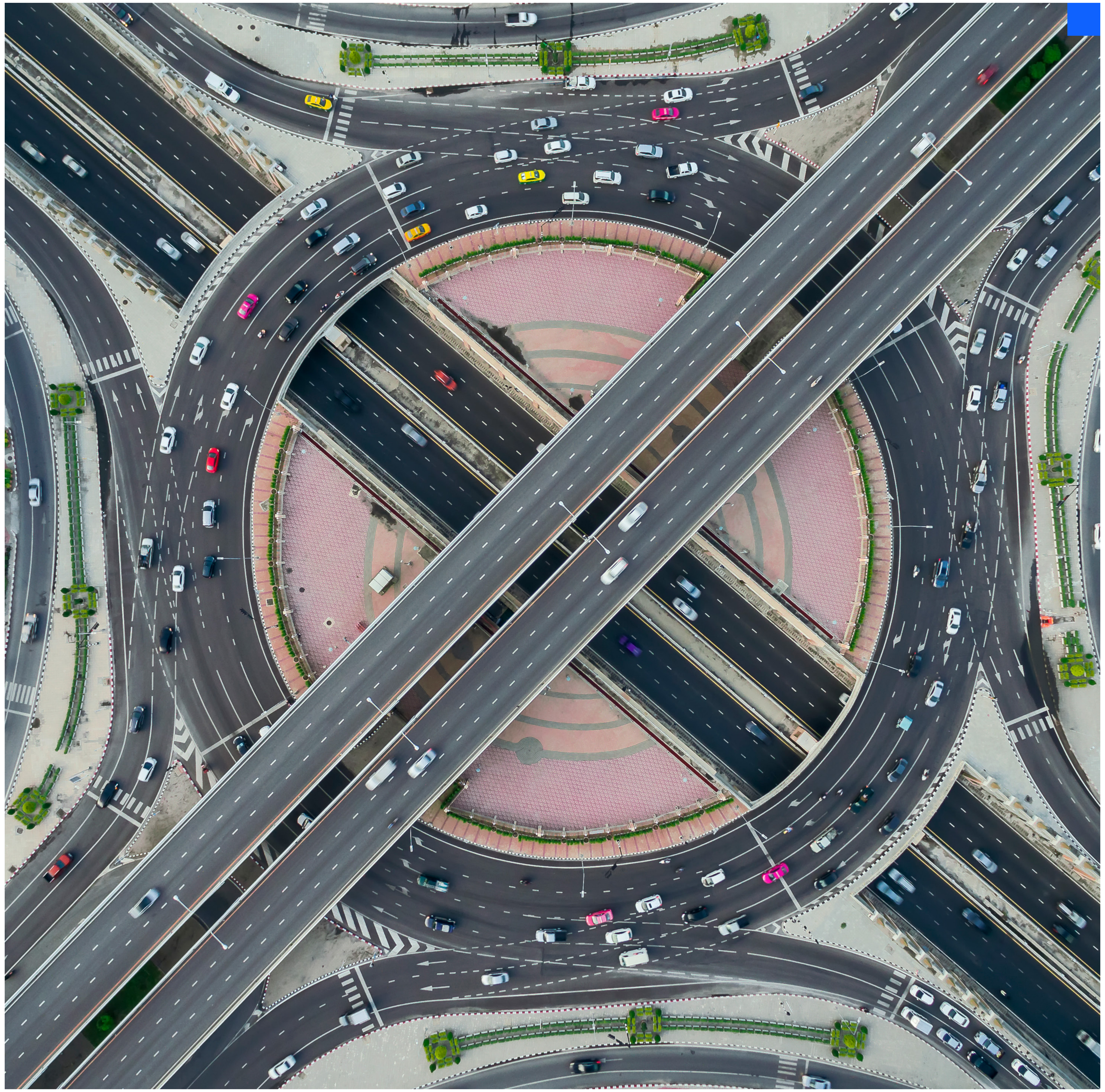
IBM® NS1 Connect

Reimagining GSLB with DNS and RUM Data



IBM®

Contents

01

Introduction

Load balancing is one of the most widely used networking technologies on the planet. Just about every networking team needs some way to optimize the use of back-end resources and ensure that no single server or environment is stretched beyond its capacity.

Global Server Load Balancing (GSLB) extends those benefits even further, ensuring the right distribution of resources between regions and hybrid environments to deliver consistent experiences wherever a user happens to be. As the adoption of hybrid and multicloud application architectures continues to expand, GSLB has changed from a niche business to a core requirement for enterprises operating at scale.

Load balancing and GSLB tools are everywhere, but do they actually contribute to business success? Most IT managers don't seem to think so. A survey by the analyst firm EMA found that most IT organizations do not currently believe their use of load balancers in the cloud is successful.

In particular, IT leaders and practitioners told EMA that "visibility suffers as [load balancing] tools struggle to collect consistent telemetry across end-to-end infrastructure. ... Larger enterprises especially struggle with visibility."

Cost is also a factor—the consensus among IT leaders and CIOs is that load balancing solutions don't deliver good value for money despite their status as critical infrastructure.

In the search for load balancing and GSLB solutions with the right feature set at a reasonable cost, many IT leaders have started to look further afield from traditional load balancers. Web Application Firewalls (WAFs), cloud-based Elastic Load Balancers (ELBs) and Application Load Balancers (ALBs), and even in-house solutions, are often used as substitutes.

In this eBook, we'll review the pros and cons of different approaches to load balancing and GSLB. We'll look at traditional inline load balancers, WAFs, ELBs/ALBs and in-house solutions to see where their strengths and weaknesses lie. Then we'll look at new, disruptive approaches that may offer the mix of functionality, resilience and cost effectiveness that many businesses are looking for.

Inline load balancers

Inline GSLB solutions all operate the same way. End-user device traffic is directed to a single ingestion endpoint and then routed to back-end resources. The load balancer uses availability data drawn from back-end resources to inform the way that it directs traffic, minimizing latency and preventing downtime.

There are two inherent challenges with this approach:

No visibility into end-user devices:
Inline GSLB solutions capture plenty of data from servers and clouds to inform load balancing decisions once an inbound query reaches the ingestion endpoint. Yet they are completely blind to everything that happens before that. This lack of end-to-end visibility allows latency to linger where it matters most—in "last mile" connections.

Single point of failure:
Directing all inbound traffic to a single GSLB endpoint creates an inherent vulnerability. If that ingestion endpoint goes down or is misconfigured, the entire back-end infrastructure becomes inaccessible. Beyond the issues created by the architecture of inline GSLB solutions, they often deliver poor value for the limited functionality they deliver. Between per appliance (physical or virtual) costs, software licenses, professional services and periodic upgrades, most large enterprises end up with eight-figure annual bills.

The reality is that GSLB is an afterthought for most load balancing companies. Their primary focus is on local load balancing use cases—that's where most of their business comes from. Since GSLB is a smaller niche, it generally doesn't get the attention it deserves, in spite of the fact that users of GSLB are some of the largest and most sophisticated enterprises in the world.

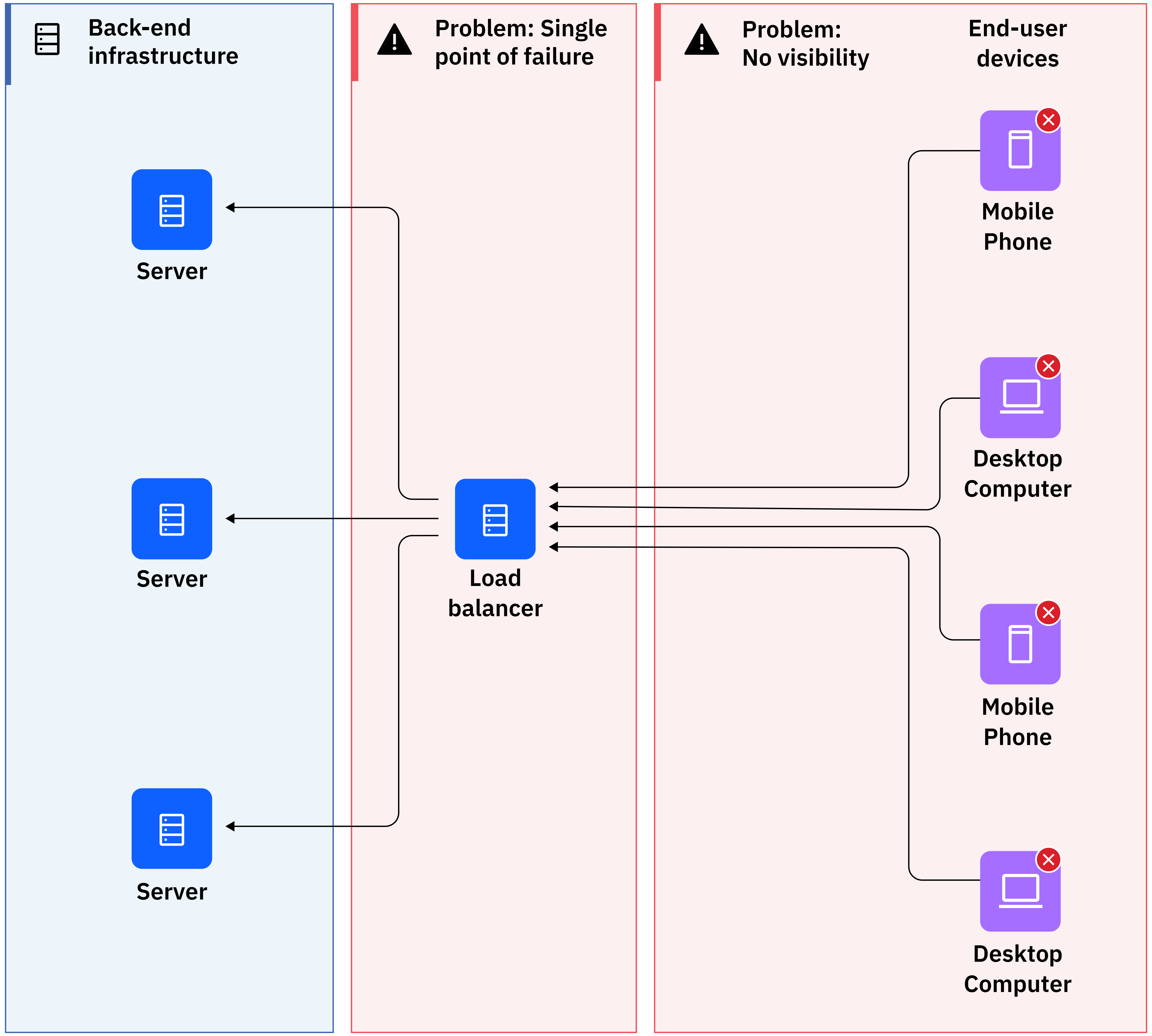


Figure 1. How inline load balancers work.

Cloud or SaaS WAF

Some companies integrate load balancing functions into their web application firewall (WAF) through an application delivery controller (ADC) configuration. This approach allows multiple functions to happen at the same network entry point, including load balancing, http/https traffic screening, DDoS protection and more.

Using ADC-based WAFs for load balancing is a popular option, but the inline placement of these solutions leads to similar issues faced by standard GSLB solutions, namely: no visibility into end-user devices and single point of failure and layer 7 only.

No visibility into end-user devices:
Like inline load balancers, WAFs are entirely focused on what happens after queries come in through the ingestion endpoint. The result is the same: increased latency in "last mile" connections.

Single point of failure:
ADC-based WAFs have the same weakness as inline load balancers —all inbound traffic comes in through a single endpoint. If that ingestion endpoint goes down or is misconfigured, the entire back-end infrastructure becomes inaccessible.

Layer 7 only:
WAFs are mainly designed to handle http and https traffic, which means that using them as a load balancing option is inherently limiting. All other protocols go unbalanced when you use a WAF for load balancing.

If GSLB is a niche market for inline load balancing companies, it's barely on the radar screen of the WAF providers. Nobody buys a WAF for its GSLB capabilities—it's almost coincidental that it works for this use case at all. For that reason, WAF providers aren't exactly tuned in to the load balancing requirements of their customer base. If it gets any attention at all, load balancing is treated as a sideshow.

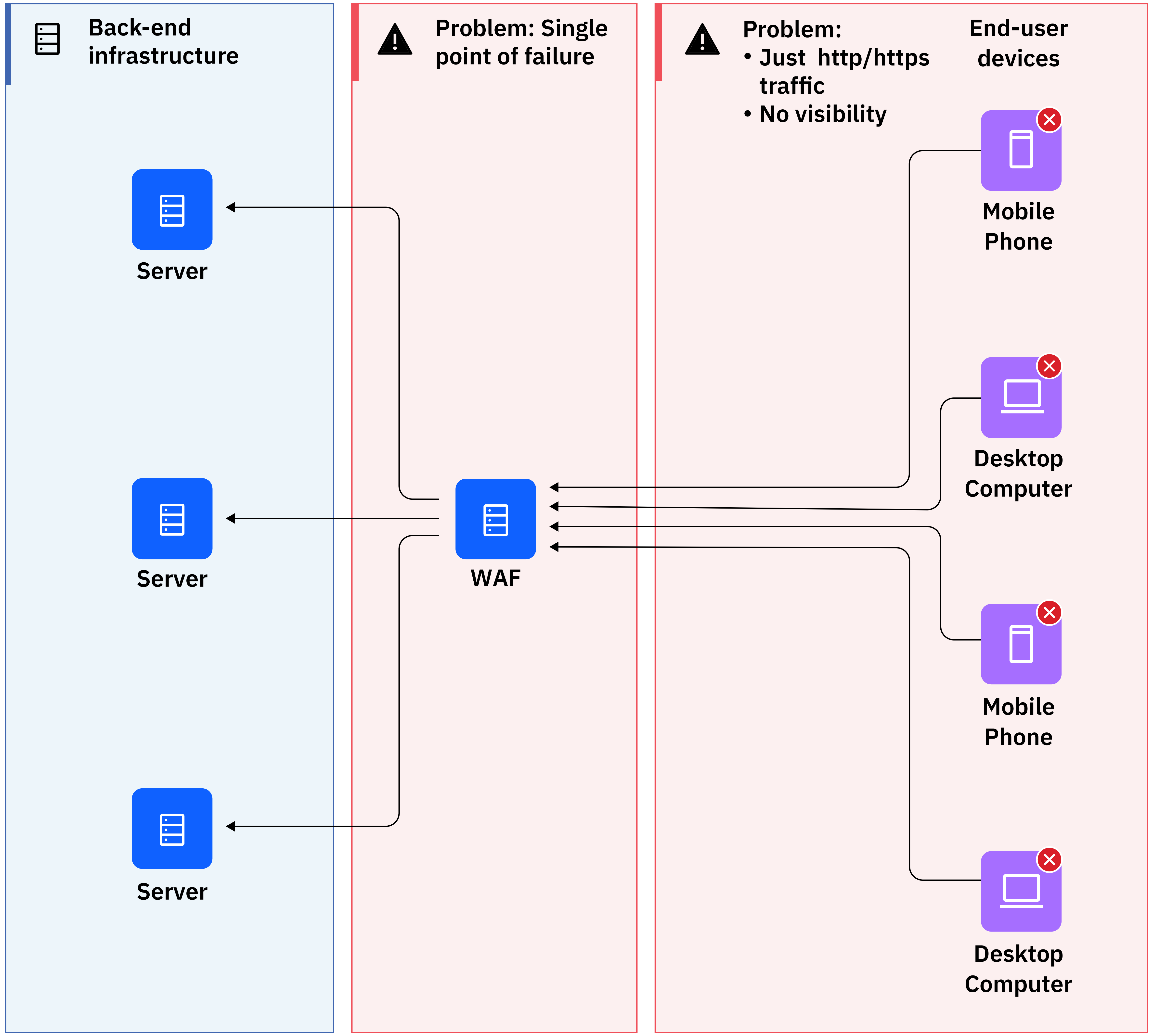


Figure 2. How WAFs work.

Cloud ELBs/ALBs

Cloud providers usually offer load balancing in the form of elastic load balancers (ELBs) or application load balancers (ALBs). These solutions rely solely on BGP weighting to balance loads —they don't have the ability to ingest latency or other internal measurements to inform a more sophisticated approach to load balancing.

If you're using a single cloud, ELBs/ALBs might appear to be the only load balancing you need. Yet there are some inherent downsides to relying solely on these solutions:

No visibility into end-user devices: Like inline load balancers and WAFs, cloud ELBs and ALBs have no visibility into "last mile" connections.

No data ingestion capability: ELBs/ALBs are limited by their BGP-based functionality. They don't have the ability to ingest data or make traffic-steering decisions based on data inputs.

No ability to balance loads between clouds: Cloud ELBs/ALBs are only designed to work within a single environment. If you're using multiple clouds, they don't offer a way to balance loads between environments.

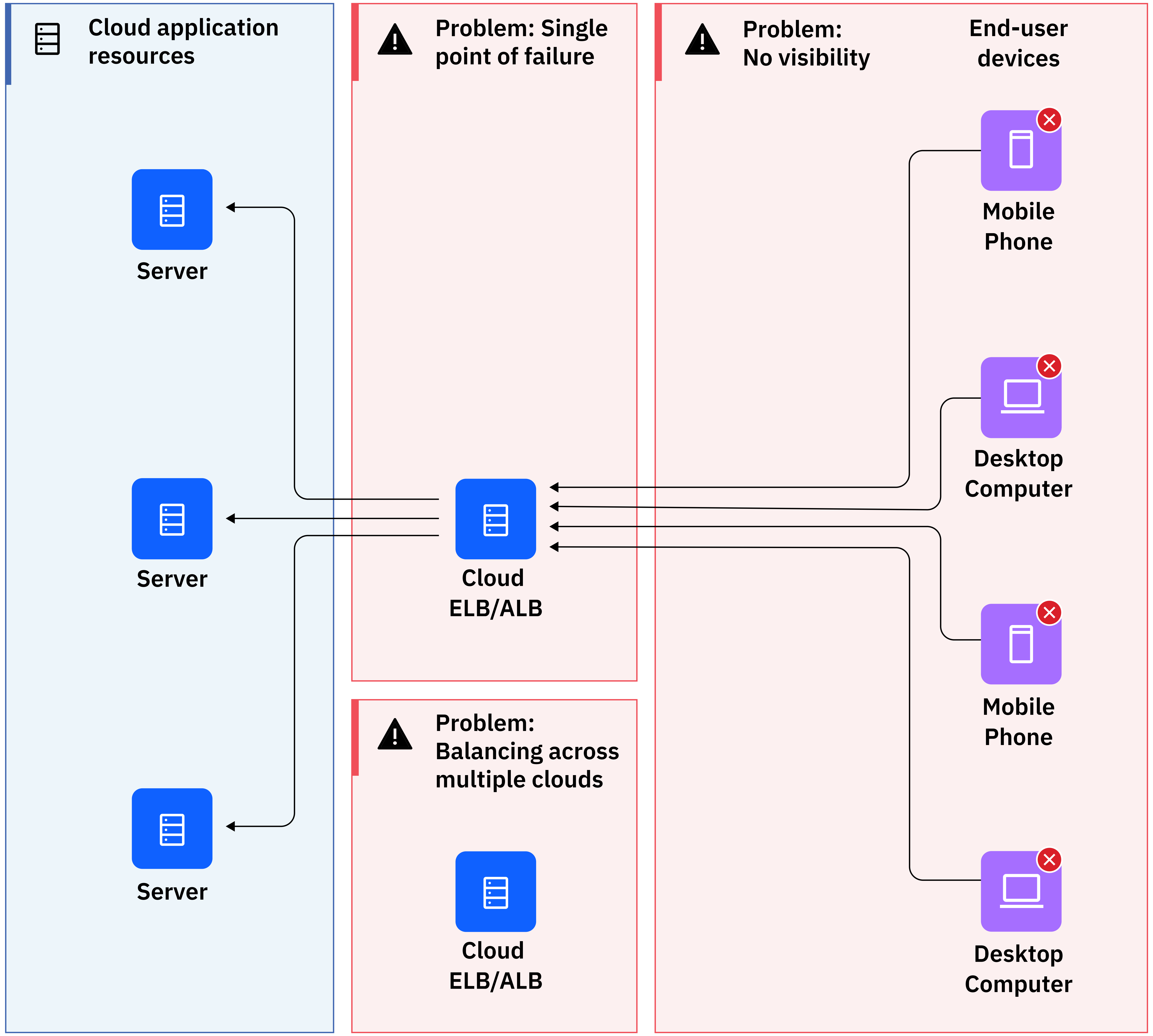


Figure 3. How cloud ELBs/ALBs work.

In-house load balancers

Given the functionality limitations and significant costs of GSLB solutions, some companies simply decide to build their own. The advantages of this approach are clear: you can customize the solution for specific business needs, avoid paying for functionality you'll never use and control the outcomes of critical issues like security.

High long-term cost:
At the same time, the long-term cost of developing, maintaining and supporting an in-house solution can be significant. Your team needs to either have the skillset to plan and develop the tool already or acquire those skills. Institutional memory and continuity planning also become important—anything developed by an in-house team needs to survive when team members depart. The team needs to be large enough to troubleshoot the solution 24/7/365—a significant challenge for global enterprises with under-resourced technology teams.

No shared risk and responsibility:
There's also the question of risk and responsibility. When you have a third-party load balancing vendor, they will naturally share the blame if something goes wrong. With in-house solutions, you have no one to blame but yourself if the solution causes an outage or disrupts business processes.

A different path: DNS and RUM data

Inline load balancers, WAFs and ELBs/ALBs all involve significant operational sacrifices. Whether it's functionality, resilience, cost, protocol coverage or some combination of these factors, most GSLB solutions simply don't fit the requirements of the large enterprises that need GSLB the most.

That's why IBM® is charting a new path —one which uses DNS and real user monitoring (RUM) data to disrupt the GSLB market by providing a low-cost, resilient, high-performance solution with end-to-end visibility.

The importance of DNS
Inline load balancers already use DNS to route traffic between servers, clouds and regions. DNS-based load balancing has a clear advantage over WAF load balancing, as it handles the complete spectrum of internet traffic, not just http/https queries. It's also superior to BGP-based load balancing because it takes latency—not simply the number of "hops"—into account when choosing a connection pathway.

DNS and load balancing
While the use of DNS gives inline load balancers an advantage over WAFs and ELBs/ALBs, the issue has always been placement—not how loads are balanced but where they are balanced. Placing load balancers in the resolution pathway with a single ingestion endpoint ignores the impact of "last mile" latency while creating a single point of failure.

Benefit of an end-to-end approach
Ideally, you want an end-to-end approach to load balancing—one that allows you to direct traffic from the moment a device generates a query all the way into the back-end resources that resolve it. This would provide control over every step of the resolution pathway without creating a bottleneck that poses the risk of downtime.

The DNS solution
This is where DNS offers an ideal solution. Authoritative DNS providers already control the resolution pathway to the "last mile"—they have both the visibility and control needed to balance loads end to end without creating a bottleneck in the resolution pathway.

> “Authoritative DNS providers already control the resolution pathway to the “last mile”—they have both the visibility and control needed to balance loads end to end without creating a bottleneck in the resolution pathway.”

RUM data draws information from both user devices and back-end services to create a highly accurate, granular, real-time picture of availability and performance.

The value of RUM data
If authoritative DNS is the best mechanism to implement end-to-end GSLB decisions, what dataset will it use to determine the appropriate pathways? Many premium authoritative solutions have traffic-steering capabilities, but the datasets they use to guide those decisions vary quite a bit.

At a basic level, performance monitors can improve load balancing by informing your DNS system about how back-end systems might impact the user experience. If a system is slow or down, DNS can steer traffic to resources that are available and performing better.

Performance monitors are only a piece of the GSLB equation, however. They can tell you how your systems contribute to user experience, but they aren’t measuring the user experience itself. Since the ultimate goal of GSLB is to improve the speed, reliability and availability of applications for the benefit of end users, it’s important to focus on the data that can help you meet that goal directly. This is where RUM data comes into the picture.

RUM data is the gold standard for measuring the quality of experience for end users. RUM data draws information from both user devices and back-end services to create a highly accurate, granular, real-time picture of availability and performance. With RUM data you can make informed decisions about how load balancing impacts user experience.

Some companies produce their own RUM data by gathering information directly from the applications they deliver around the world. If the sample size is large enough and the data is gathered on a regular enough cadence, it forms an accurate picture of user experiences for relevant regions and device types.

Most companies lack the mechanism to collect RUM data, the global scale to build a representative sample, or both. To balance loads appropriately across regions and device types through DNS, they need someone else to collect and process the data on their behalf.

NS1 Connect has the largest selection of traffic-steering filters, and its unique filter chain technology allows you to customize your traffic-steering logic through the use of multiple filters in tandem.

Introducing IBM NS1 Connect GSLB

IBM now offers a new GSLB solution that leverages DNS and RUM data to balance traffic while providing superior resilience, lower connection latency and lower operating costs. The IBM® NS1 Connect solution.

The IBM® NS1 Connect solution for authoritative DNS already contains one of the most sophisticated, customizable traffic-steering offerings on the market. NS1 Connect has the largest selection of traffic-steering filters, and its unique filter chain technology allows you to customize your traffic-steering logic through the use of multiple filters in tandem.

The robust NS1 Connect RUM dataset distinguishes it from other authoritative DNS providers and provides the engine for context-driven GSLB decisioning at scale. NS1 Connect RUM data comes from a global network of users and devices representing the most common ISPs in every region.

RUM data is continuously updated, so you know exactly how the GSLB decisions you make impact user experience. Have your own RUM dataset you'd like to use? No problem! With IBM® NS1 Connect you have the choice of using our native RUM data or ingesting your own RUM data into the platform. Either way, you'll get the real-time performance data you need to guide GSLB decisions through DNS.

As an out-of-band solution, IBM® NS1 Connect makes load balancing decisions without the need to insert a single point of failure into the resolution chain. Queries are seamlessly directed from the user's device all the way to application workloads along a pathway that avoids slow or unavailable resources.

That same out-of-band architecture, when combined with RUM data, also provides visibility and control over "last mile" connections that often introduce latency and degrade application performance.

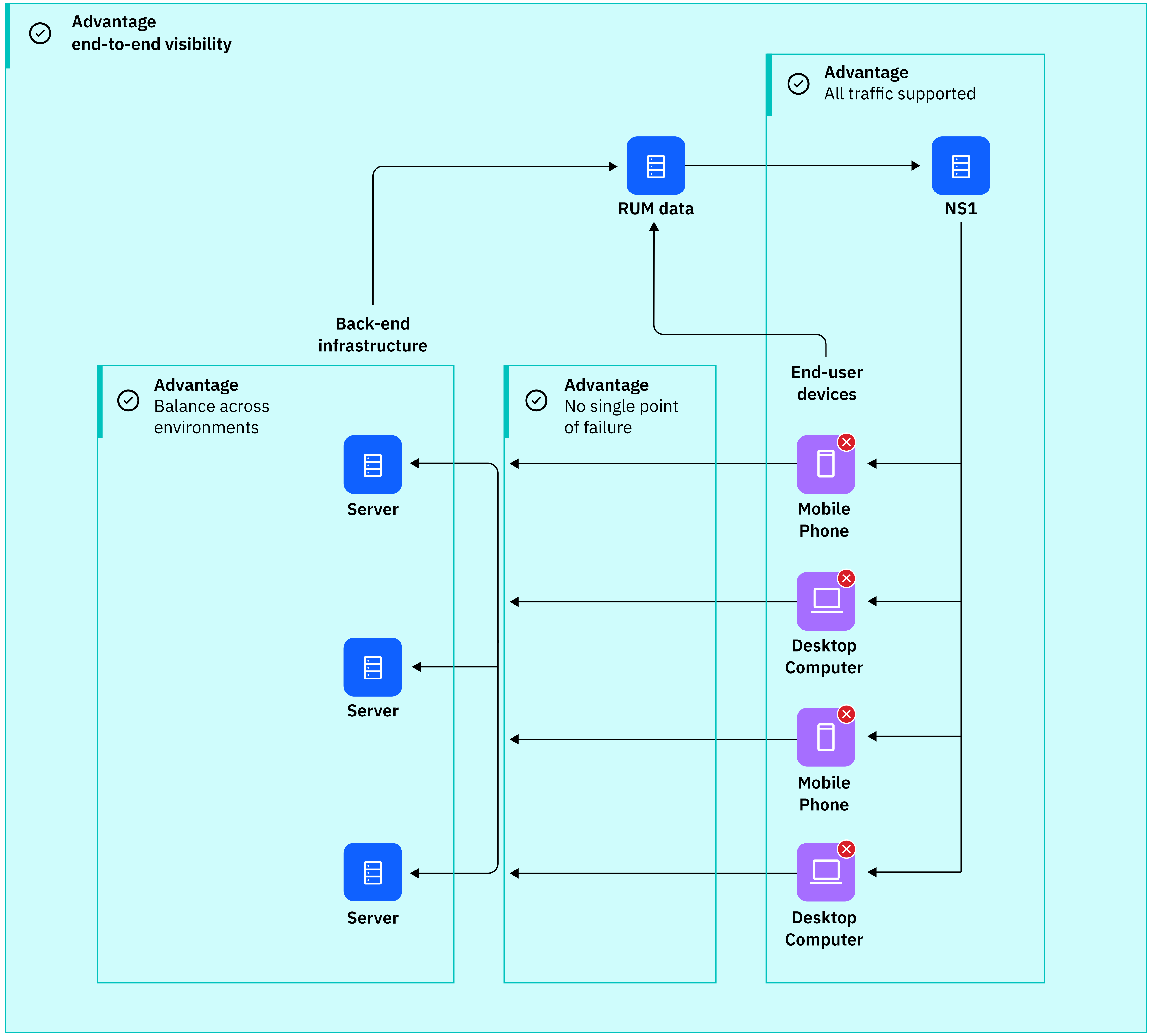
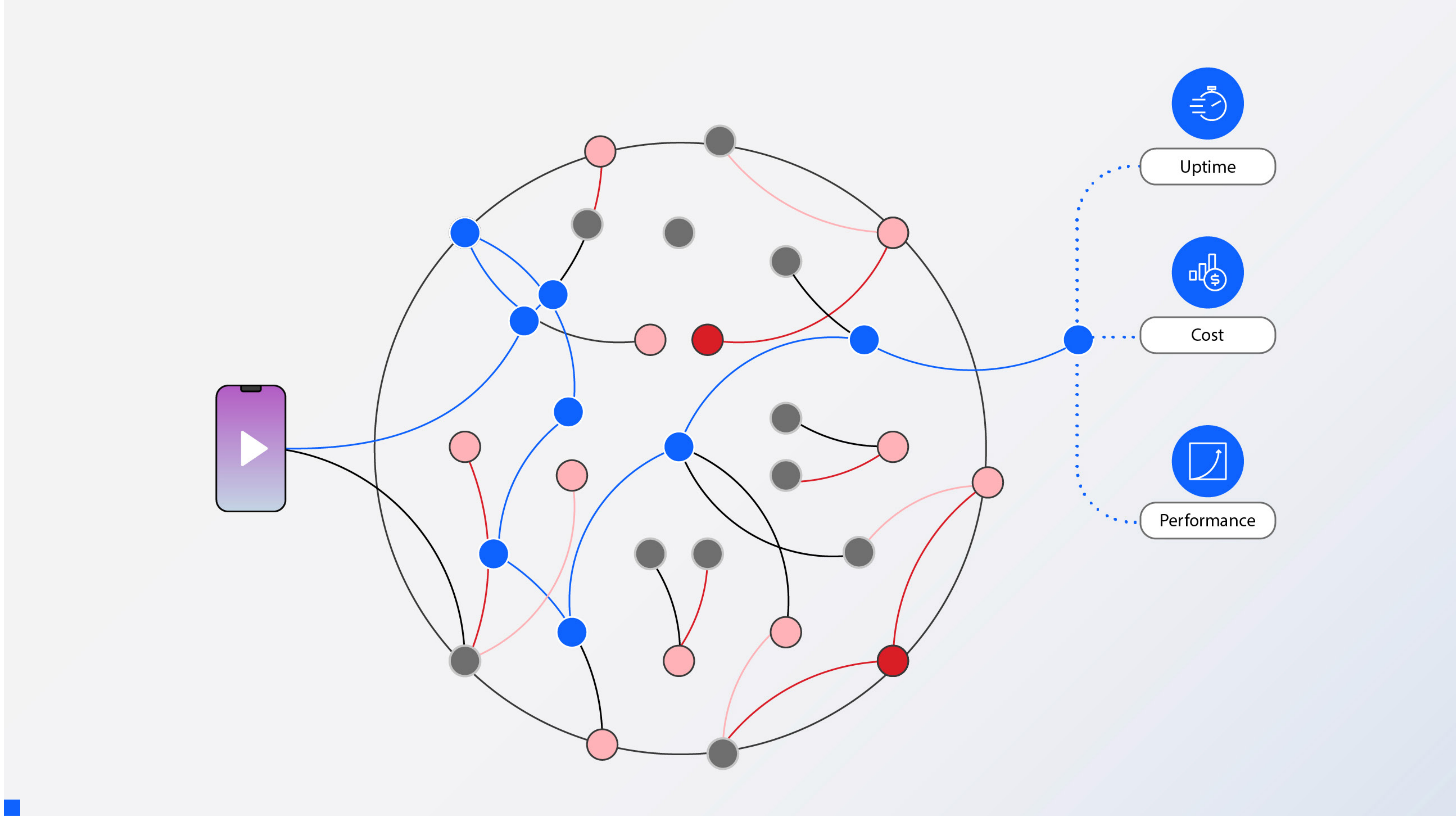


Figure 4. How NS1 Connect GSLB works.

Where inline GSLB solutions only have an impact on traffic once it enters the “front door” of a company’s infrastructure, out-of-band DNS-based GSLB solutions have an impact from the moment a user’s device generates a query. This end-to-end approach pays major dividends by lowering latency and improving application performance.

Cost is consistently cited as one of the most prominent issues with existing load balancing solutions. Between the appliances and licenses of inline load balancers and the overwhelming complexity of ELB/ALB pricing, it’s no wonder that many IT managers are looking for a better way. As a SaaS product delivered without the operational overhead of legacy load balancer architectures, IBM® NS1 Connect is offered at a far lower price point than competitive solutions.



08

Conclusion

Leveraging the power of DNS

At a basic level, DNS measures intent. Whenever a user opens an app, types in a URL or clicks on a link, they're offering the first tangible expression of what they want an application to do. Load balancing at that first expression of intent is the most efficient and effective way to deliver quality user experiences. By the time a query reaches an inline load balancer, WAF, cloud environment or in-house load balancing solution, a significant portion of the recursion chain has already passed by. The opportunity to optimize connection speed and minimize latency is limited.

By leveraging the full end-to-end journey, DNS offers the ability to improve application performance at every stage of the recursion chain. The addition of RUM data adds necessary context and granular insight to find the best available connection for every device at any given moment in time.

If you're ready to experience the power of DNS and RUM data, looking to lower your load balancing costs, or struggling with end-to-end visibility in your application, NS1 Connect deserves consideration.

Learn more about NS1 Connect GSLB →

IBM Corporation
New Orchard Road
Armonk, NY 10504

Produced in the United States of America
January 2024