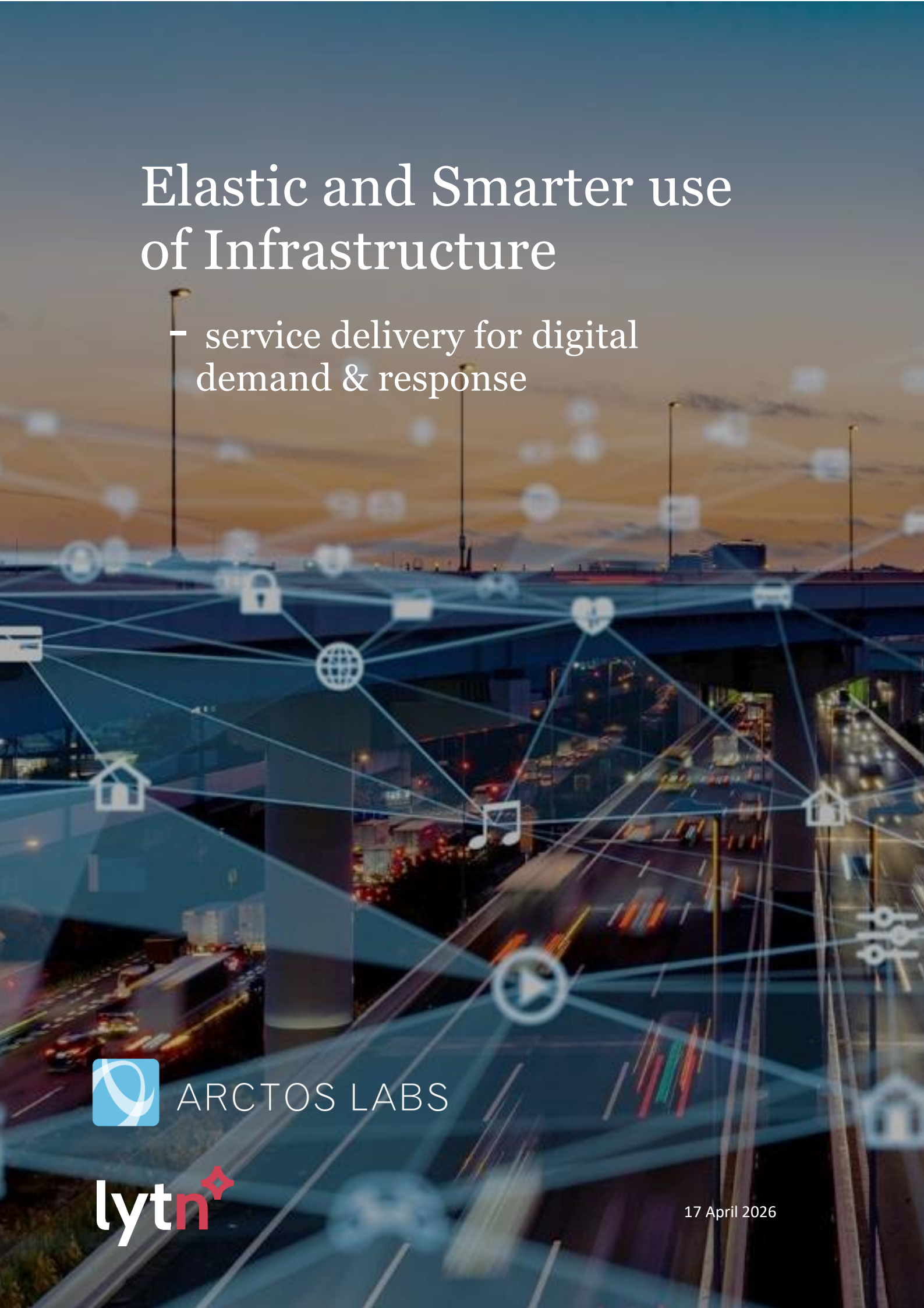
Elastic and Smarter use of Infrastructure

– service delivery for digital demand & response



ARCTOS LABS

lytn

17 April 2026

1 Introduction

Modern IT infrastructure strategies are being reshaped by a couple of important trends:

- Applications are increasingly moving closer to end users to meet stringent Quality of Experience (QoE) requirements
- Explosion of data usage calls for efficient data platforms and considerations on how data flows in the infrastructure
- AI-driven applications are entering the scene, adding to compute and data consumption, and driving the issue of their locality
- Applications are experiencing increasing variations in intensity, every business day is different
- Enterprises are placing greater scrutiny on infrastructure costs across both cloud and on-premises environments
- Environmental concerns and new regulations are pushing data centre efficiency and energy consumption
- Data sovereignty and regulatory requirements are adding new layers of complexity

Individually, each of these trends introduces operational challenges for enterprises directly and for their service providers. Collectively, they make it significantly more challenging to achieve **cost-efficient, high-performing infrastructure operations** that support the needs of the applications running on them. In short, aligning response to the demand also taking constraints at any point of time becomes challenging, every day.

This paper explores two complementary mechanisms to address these circumstances:

- **Elastic and Smart use of infrastructure,** enabling enterprises and service providers to utilize existing infrastructure more efficiently by controlling resource allocation, or shifting workloads to locations to best meet the intended cost and business targets. This is achieved by smart, in-time optimization based on prediction and business targets.
- **Infrastructure capacity analysis** enables enterprises and service providers to adjust installed capacity and decide locations of their infrastructure to balance cost, performance, and service quality. Such analysis helps enterprises and service providers to evolve their infrastructure to meet QoE and cost targets, including a balanced approach in using public clouds.

2 Baseline: Static, Planned Deployment

Traditionally, infrastructure planning follows a static model. IT administrators analyze application requirements and determine at which location applications should be deployed. Such analysis is made up from several assumptions, including distribution of users in geography, usage intensity, cost of compute (Capex and Opex), needed application QoE, etc.

Author: Mats Eriksson Date: 27 April 2026

During the lifetime of applications, such static approach makes it difficult to use emerging providers and locations that could be candidates for deployment, often leading to an unnecessary conservative approach.

While the application deployment may be automated using a DevOps approach with various orchestration tools, the decision-making process on locality and resource allocation is largely manual and, at best, revisited at periodic intervals.

Finally, it is difficult in most cases to check the actual values of the network Service Level Objectives that are part of the SLA agreed upon with Service Providers.

This typically results in:

- A static placement of workloads across private data centers and cloud providers, where workloads with dynamic loads often are deployed in the cloud, or resources over-provisioned
- Over- or under-utilization of private / on-premises resources resulting in either QoE issues or unnecessary costs
- Over-reservation of cloud resources results in unnecessary cloud spend

The lack of **holistic and dynamic control** of how such a distributed and heterogeneous infrastructure is utilized inevitably leads to higher costs and potentially application QoE issues. As enterprises become increasingly depending on their IT infrastructure, the need for an **elastic, balanced and optimized usage** becomes a key enabler for achieving business targets.

3 Elastic and Smart use of Infrastructure

A more advanced approach introduces a new governance layer that oversees and controls the set of infrastructure at hand for the enterprise or service provider. Such a governance layer needs to be capable of reacting to real or predicted application demand to achieve an elastic and smart response.

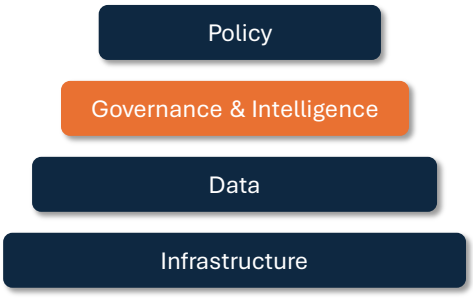


Figure 1: The emergence of a new layer for governance & intelligence

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152
Lytn SAA, 450 RUE B. powell, 34000 MONTPELLIER, France. Reg 881850465 Lytn.io. Contact@lytn.io

At first, we need to introduce a **capability that dynamically can control resource allocation and dynamically placing workloads** based on real-time demand conditions.

At its most basic level, this capability involves:

- Monitoring utilization across candidate environments
- Evaluating cost differences between locations
- Evaluating the network Service Level delivered by candidate locations
- Selecting the most efficient resource allocation and placement for each workload regularly

Additional capabilities needed include also:

- Control of infrastructure spanning public cloud and private data centres
- Taking latency constraints for applications into account
- Ability to balance workload placement taking locality of data into account
- Complying with security, and data sovereignty requirements

…just to name a few.

A dynamic workload placement capability addresses the need to place workloads sufficiently close to where its users are, to achieve the needed QoE, but also balance in cost, data transport, and constraints related to sovereignty. An optimized application delivery may also include taking energy consumption into the equation for both cost and sustainability reasons.

An effective solution must therefore continuously balance these factors, thereby enabling an elastic and smart infrastructure.

Arctos Labs ECO addresses this capability by providing smart analysis and dynamic recommendations for resource allocation and workload placement. By continuously evaluating infrastructure conditions and constraints, it enables enterprises and service providers to move beyond static planning toward **adaptive and optimized infrastructure**.

Dynamic resource control and workload placement becomes significantly more powerful when combined with the capability to **predict workload demand and network communications needs, across such a distributed infrastructure**.

Understanding how workload demand will evolve over time allows the dynamic control to:

- Scale capacity just in time.
- Proactively revisit placement decisions and migrate workloads.
- Avoid performance bottlenecks before they occur.
- Maintain consistent application quality by migrating workloads or changing resource allocations.

Author: Mats Eriksson Date: 27 April 2026

- Balance business targets vs costs during a period of high spikes, where a lower QoE might be justified

Predictive models can forecast workload demand over the coming time periods, enabling more informed and forward-looking decisions.

Lytn provides advanced prediction technology capable of delivering high-precision forecasts of workload behaviour. By integrating such insights, enterprises and service providers can shift from reactive to **proactive infrastructure control**.

The governance & intelligence layer described by this paper integrates to a suitable infrastructure orchestration capability. This orchestration translates recommendations and decisions into the appropriate concrete actions towards the infrastructure.

These actions include to reserve and allocate infrastructure capacity, to deploy and activate application software, to connect deployed software, and a range of other actions.

The dynamic workload placement and prediction described by this document are agnostic to the implementation of the orchestration. It means enterprises and service providers may use Kubernetes and OpenStack from a range of different vendors, or use OpenNebula, or VMware, for example.

4 Infrastructure Capacity Analysis

In addition to elastic and smart infrastructure, long-term efficiency requires a deeper understanding of how infrastructure is utilized over time. Such analysis will determine the justification of investment in new infrastructure, its dimensioning, and distribution geographically.

Such analysis starts by collecting and analysing:

- Usage metrics for the infrastructure acquired
- Application performance KPIs
- Cost information for the usage of cloud or private infrastructure as well as for any candidate cloud providers
- Profiling information for applications infrastructure resource needs
- Financial information for private infrastructure

From this data, enterprises and service providers can derive actionable insights, such as:

- **Whether infrastructure capacity should be redistributed across locations** to create a better match between expected usage demand distribution and infrastructure capacity distribution.
- **Which capacity should be supplied from the Edge, the Cloud**, and which cloud regions? Cloud is inherently elastic, but that comes at a cost as well as implications for QoS.

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152
Lytn SAA, 450 RUE B. powell, 34000 MONTPELLIER, France. Reg 881850465 Lytn.io. Contact@lytn.io

- **Which level of reservation is suitable for cloud capacity**? For the capacity from cloud that is stable, it makes sense to utilize resource reservation schemes for lower costs, but that is a complicated analysis.
- **Where capacity can be reduced without seriously impacting service quality**. When utilizing the ability to move workloads horizontally, additional capacity savings could be unlocked.
- **Where additional capacity could improve application performance.** In the case of dynamic workload placement, some workloads could have been placed in second-best locations still with acceptable QoS. This may imply that capacity should be increased in some locations or moved between locations.
- When is the infrastructure **upgrade to a newer generation optimally scheduled**? The financial efficiency of an infrastructure depends on many parameters, such as depreciation, power consumption, and compute efficiency. These factors impact the optimal time for upgrading the infrastructure to newer generations.

Importantly, this analysis does not only focus on cost reduction—it also considers the potential impact on application performance and user experience by understanding the capability to re-distribute load horizontally to make the most out of available infrastructure whilst controlling application QoE.

These optimization models are encapsulated within **Arctos Labs ECO**, enabling enterprises and service providers to make **data-driven decisions about infrastructure strategy, including powerful what-if analysis** rather than relying on static assumptions.

5 Start your journey

If the content of this paper resonates with your needs, we would like to explore how Arctos Labs and Lytn could enable your infrastructure to be more elastic and dynamic.

Please reach out to:

Mats Eriksson, Arctos Labs, mats.eriksson@arctoslabs.com, www.arctoslabs.com

Etienne Coulon, Lytn, etienne.coulon@lytn.io, www.lytn.io

Exhibit A: Evolution of Quality in Network Infrastructures

Dimension	**QoS** Quality of Service	**QoE** Quality of Experience	**PEI** Proactive Experience Index
Core Question	How well is the network performing?	Was the user experience impacted?	Will the experience be at risk?
Primary Focus	Infrastructure performance (latency, loss, throughput)	User impact traceability and accountability	Predicting and preventing future degradation, be actionable
Time Orientation	Present (real-time metrics)	Past (after the experience occurred)	Future (short-horizon predictions)
Operational Mode	Reactive troubleshooting	Reactive validation of impact	Proactive prevention and early intervention
Value for the Business	Ensures technical compliance	Ensures impact traceability and accountability	Ensures continuity, efficiency, and resilience

Author: Mats Eriksson Date: 27 April 2026