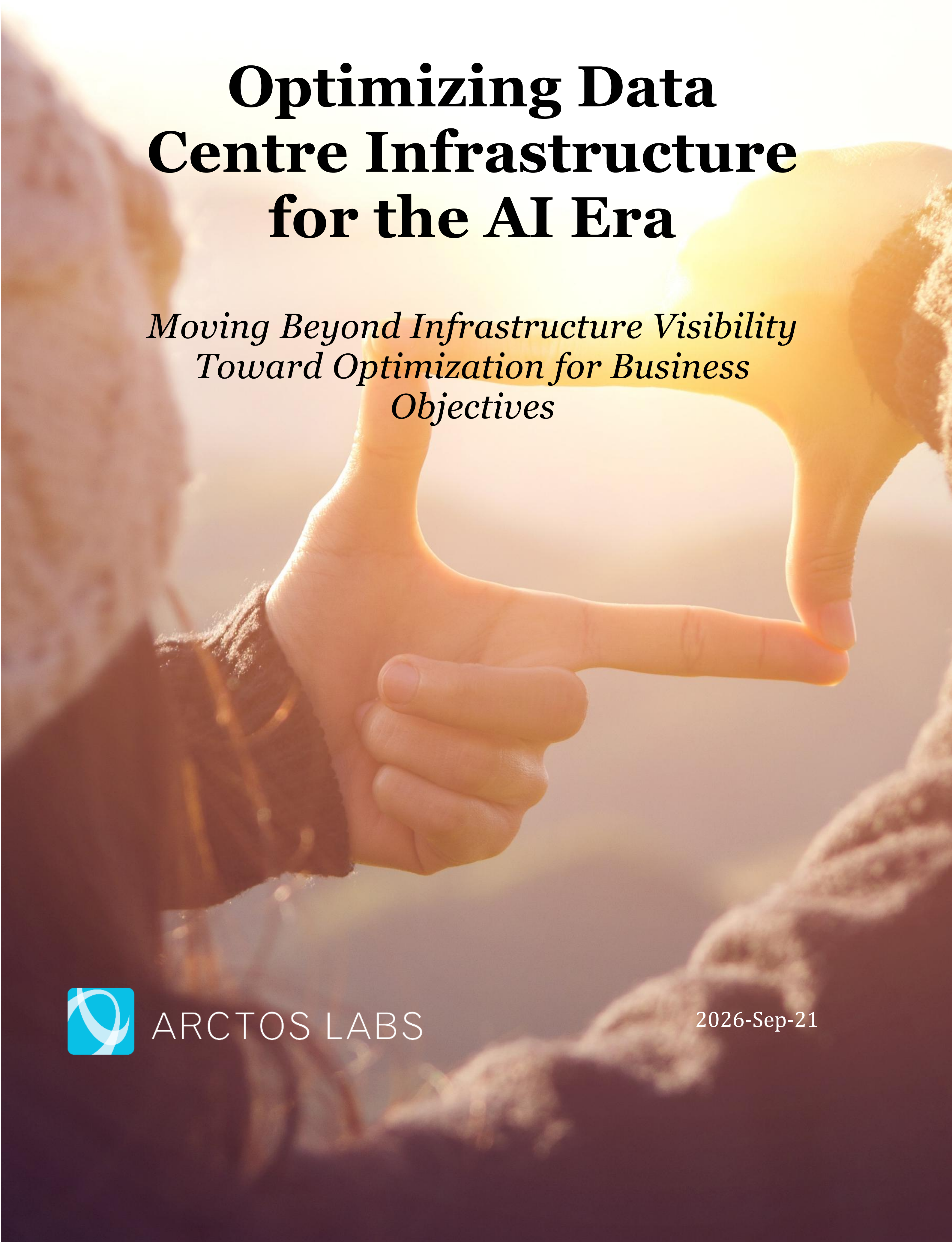

Optimizing Data Centre Infrastructure for the AI Era

Moving Beyond Infrastructure Visibility Toward Optimization for Business Objectives

ARCTOS LABS

2026-Sep-21

1 Executive Summary

Data centres are entering a period of fundamental change.

The growth of AI and other compute-intensive applications is driving higher infrastructure densities, and placing new demands on power, cooling, networking and physical space. At the same time, data centre environments are becoming increasingly heterogeneous, combining different generations of servers, accelerators, storage systems, networking technologies, cooling architectures and distributed facilities.

These developments are changing the nature of data centre planning.

Historically, capacity planning could often be approached by considering resources individually: available rack space, power capacity, cooling capacity, or compute capacity. Increasingly, these resources are interdependent. A rack may have physical space available but insufficient power. A facility may have sufficient power but inadequate cooling. A workload may have access to compute capacity but fail to meet requirements for latency, availability, data locality or cost.

The basic planning question is therefore changing:

Figure 1: Towards Focus on Business Objectives

Data Centre Infrastructure Management (DCIM) provides an important foundation for answering this question. Modern DCIM platforms combine infrastructure inventory, capacity information, power and environmental data, planning and operational workflows. Some already provide optimization and AI-assisted placement capabilities.

The opportunity is to build on this foundation with a broader **data centre optimization layer**.

Author: Mats Eriksson (assisted by AI) Date: 22 September 2026

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

Such a layer combines infrastructure information with workload requirements, operational constraints, economic considerations, and business objectives. It can evaluate alternative configurations and determine how infrastructure should be placed, expanded, upgraded, or utilized.

The resulting evolution can be summarized as:

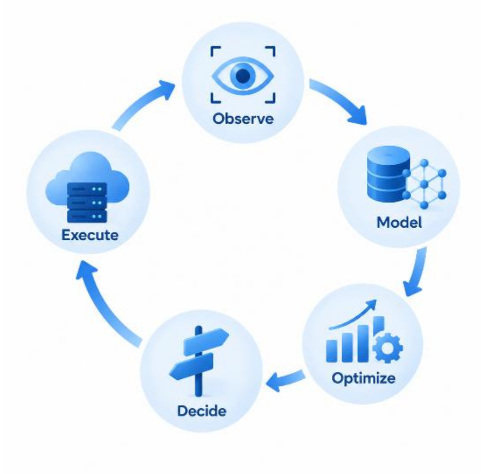

Figure 2: Evolved Optimization Loop

This paper explores what this optimization layer could look like, which problems it can address, and how it can become part of a continuous operating model for modern data centres.

---

2 The Data Centre Is Becoming an Optimization Problem

Data centre infrastructure has always involved trade-offs. What is changing is the number and interdependence of the variables involved.

A traditional enterprise data centre might have been designed around relatively predictable server capacity and relatively stable workloads. Modern environments increasingly contain high-density IT infrastructure (for AI), heterogeneous compute platforms and dynamic workloads. It means the installed compute capacity per square meter are substantially higher than previously. The types of servers installed contain different generations of silicon, accelerators, and storage systems. Workloads exhibit an intensity that are more dynamic and varying.

Power and cooling are particularly important aspects that are impacted by this change.

A new high-intensity workload (e.g. AI) can affect not only the number of servers that can be installed, but also the power distribution architecture, cooling requirements, rack placement and the remaining capacity available to neighbouring equipment.

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

Similarly, a decision to place workloads in a particular area of the data centre can affect network utilization, latency, power consumption and the ability to maintain redundancy.

2.1 Capacity is no Longer One-dimensional

Consider a facility with available floor space.

At first glance, this suggests that additional IT infrastructure can be deployed. But if the available space is in an area without sufficient power distribution, the capacity may not be usable.

Nominal capacity is not necessarily usable capacity

The same can apply to cooling, network connectivity, physical load limits, or redundancy requirements.

The practical capacity of the facility is therefore determined by the interaction between several resources and constraints.

This leads to an important distinction:

Nominal capacity is not necessarily usable capacity

Modern capacity planning therefore needs to consider not only how much compute capacity exists, but where it exists, what constraints apply to it and which combinations of resources can be used simultaneously.

IT infrastructure for AI makes these relationships particularly visible.

Artificial Intelligence increases the complexity

Such high-density infrastructure (e.g. GPU systems) can create substantially different requirements from conventional CPU-based infrastructure. Cooling architecture, power distribution, network connectivity, and rack configuration can all become important factors in determining where such capacity can be deployed.

Research is also increasingly exploring optimization approaches for data centre cooling and energy management, illustrating the growing importance of multi-variable optimization in data centre infrastructure.

AI is therefore not the only reason to rethink data centre planning. It is, however, an important catalyst.

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

3 From DCIM to the Optimization Layer

DCIM (Data Centre Infrastructure Management) has evolved considerably from its early role as an inventory and visualization system.

Modern DCIM platforms can provide a unified view of assets, racks, power, cooling, environmental conditions, and capacity. They can support capacity planning, scenario modelling, workflows and operational decision-making.

This makes DCIM an essential component of a modern optimization architecture.

But there is an important distinction between **knowing what exists** and **determining how to evolve.**

Capability	Primary question
Monitoring	*What is happening?*
Infrastructure Management	*What infrastructure do we have?*
Optimization	*What should we do?*
Planning & Execution	*How should we make it happen?*

These capabilities are complementary rather than mutually exclusive.

- A DCIM system can provide the infrastructure model.
- An orchestration system can provide the mechanism for changing IT infrastructure.
- A cloud management platform can provide information about cloud resources and costs.
- A workload platform can provide application requirements.
- An optimization layer can bring these sources together and determine which option best satisfies the defined objectives and constraints.

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

3.1 The Optimization Layer

At its simplest, the concept can be represented as:

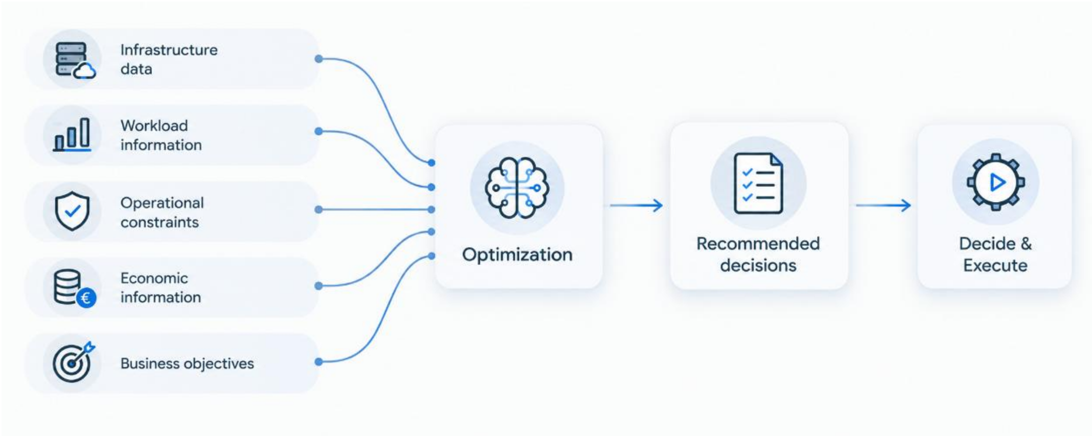

Figure 3: An Evolved Optimization Layer

The optimization layer therefore does not replace DCIM or orchestration.

Instead, it connects to them.

---

4 What Can Be Optimized?

The range of potential optimization problems is broad. However, most can be grouped into three interconnected areas.

4.1 Physical Infrastructure Optimization

The first category concerns the configuration of the physical data centre.

Examples include:

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

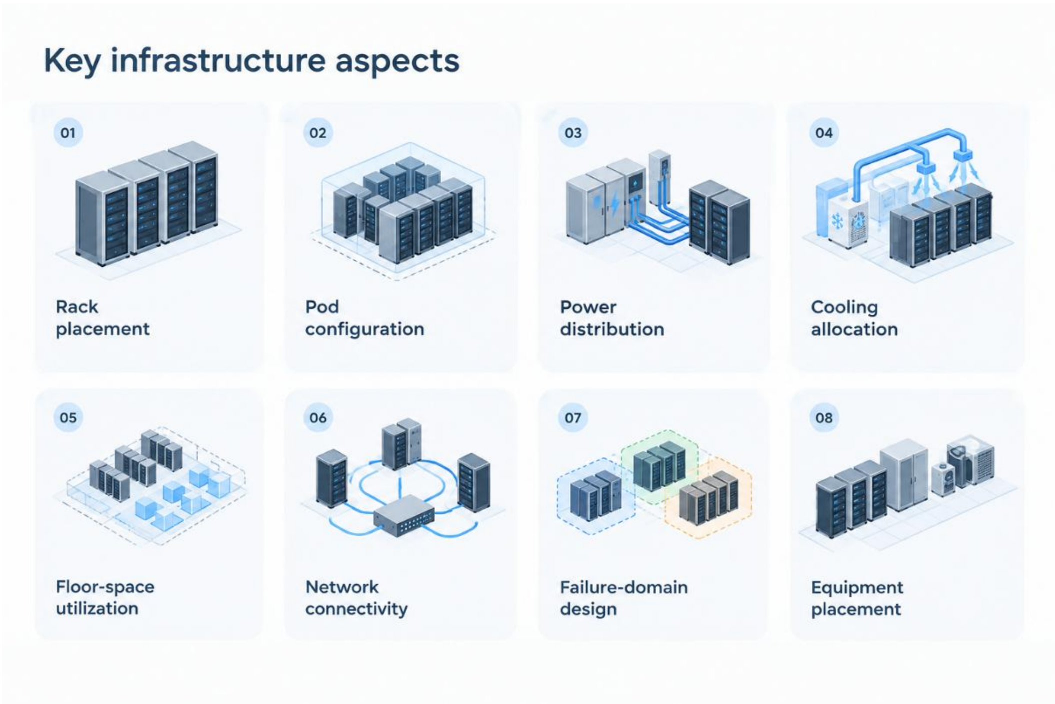

Figure 4: Aspects for Physical Infrastructure Optimization

The basic question is:

How should the existing physical infrastructure be configured?

A rack placement decision, for example, may need to consider available space, power capacity, cooling characteristics, network connectivity, physical constraints, and redundancy.

This can become difficult to solve manually as the number of racks and constraints increases.

The optimization problem is not simply to find ***a*** feasible space.

It is to find a configuration that best satisfies the objectives of the operator

It is to find a configuration that best satisfies the objectives of the operator

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

4.2 Capacity and Investment Optimization

The second category concerns how infrastructure should evolve.

These decisions have both technical and economic consequences.

For example, adding infrastructure immediately may provide additional capacity but increase capital expenditure and leave resources underutilized.

Delaying investment may improve utilization but increase the risk of capacity shortages.

Moving workloads to cloud infrastructure may avoid an infrastructure investment but introduce recurring costs, data transfer costs, or performance limitations.

The optimization problem therefore becomes:

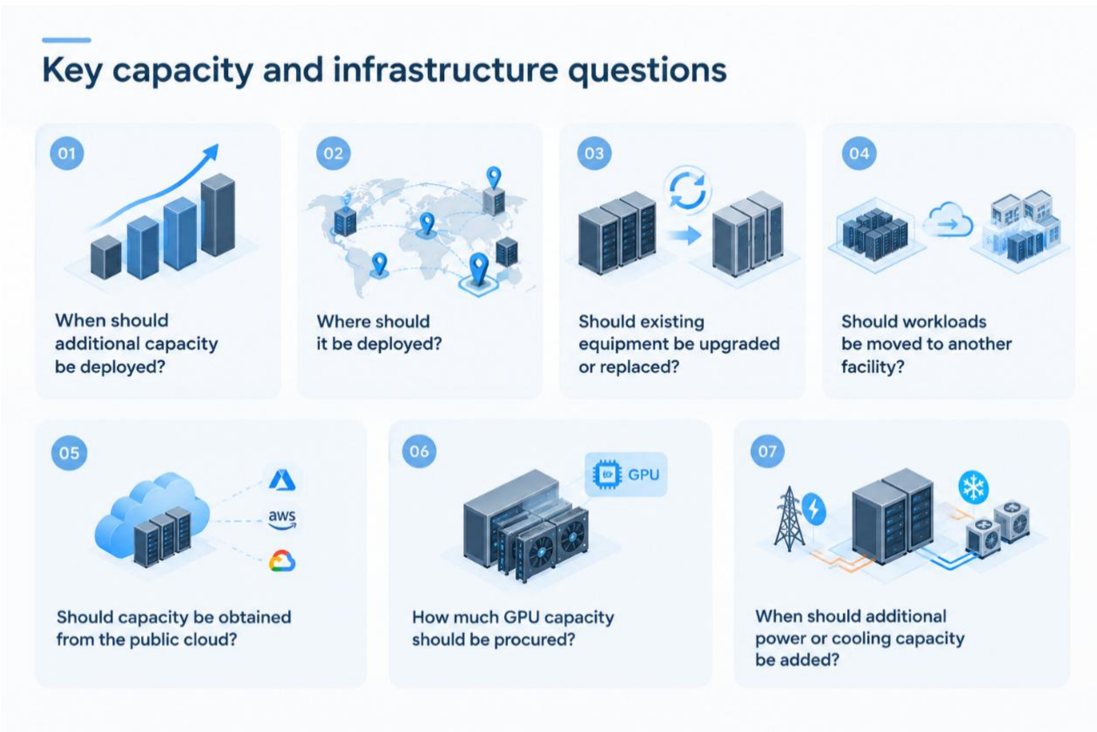

Figure 5: Key Questions for Capacity and Investments

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

4.3 Workload Optimization

The third category concerns the allocation of workloads to IT infrastructure.

A workload may have requirements relating to:

Figure 6: Workload Optimization Requirements

The question becomes:

Where should a workload run?

For a simple environment, this may be handled by rules, or manually.

For a complex distributed environment, the decision can depend on many interacting variables.

Where should a workload run?

A workload could potentially run in several locations, but those locations may have different costs, resource availability, network characteristics, and resilience properties.

Optimization can evaluate these alternatives systematically.

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

5 The Optimization Problem

A useful way to understand the optimization layer is to consider the information it can use to make a decision.

Figure 7: Holistic Optimization Model

These inputs can be represented as a common model and evaluated by an optimization engine.

5.1 Optimization for Multiple Objectives

Real data centre decisions rarely have a single objective.

Consider a rack placement decision.

- There might be rows that have available power, but limiting future POD placement and business flexibility if they are used
- There might be available spaces, but they might lead to stranded power.
- Available rows may not provide the lowest latency.

The technically optimal solution may therefore not be the economically optimal solution.

Optimization provides a framework for making these trade-offs explicit.

Instead of asking:

> **"Where can we put this rack?"**

the operator needs to explore:

> **"Which placement provides the best balance between cost, performance, capacity and resilience?"**

"Which placement provides the best balance between cost, performance, capacity and resilience?"

This distinction becomes increasingly important as data centre environments become more constrained.

---

6 From Static Planning to Continuous Optimization

Traditional data centre planning is generally periodic.

A team develops a capacity plan, makes infrastructure investments, deploys racks, and then revisits the plan as conditions change.

This can be represented as:

> **Plan → Build → Operate → Re-plan**

The environment, however, does not remain static.

- Workloads change.
- Infrastructure ages.
- Energy prices change.
- New capacity becomes available.
- Equipment fails.
- AI workloads are introduced.
- Business priorities change.

A more dynamic model is therefore possible:

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

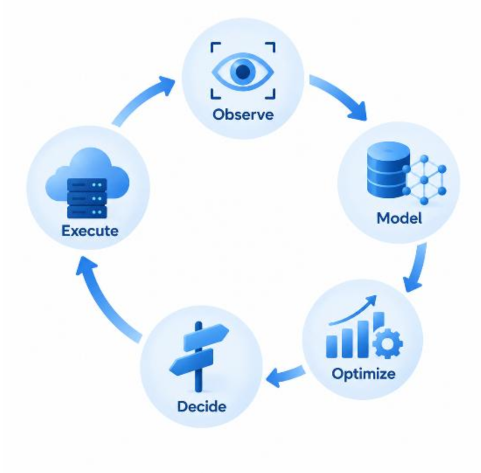

Figure 8: Evolved Continuous Optimization Loop

This creates a continuous optimization loop.

---

7 Example: Expanding AI Capacity

Consider a data centre operator that needs to expand capacity with new high-density IT infrastructure (e.g. GPU cluster) to serve growing AI business.

The operator has an operational facility with:

- Existing IT infrastructure
- Existing workloads
- Limited power capacity
- Limited cooling capacity
- Available but fragmented rack space
- Existing network infrastructure
- Defined redundancy requirements

As this is an expansion that is expected to happen gradually over time, the decision has multiple aspects balancing the short-term actions with longer term opportunities.

Planning need therefore to explore several interconnected questions.

- Where could new GPU capacity be placed? And how much?
- Can the existing power distribution support it? When would an upgrade be needed?

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

- Will the cooling system support the resulting thermal load, or is an upgrade needed?
- Which placements create a localized constraint, with stranded resources?
- Does the network architecture provide the required connectivity? Or do we need an upgrade?
- Should existing workloads be re-balanced to other areas in the facility to create the needed headroom for the new GPU capacity?
- Would a different configuration leave more capacity available for future expansion?
- Would upgrading existing infrastructure be more attractive than deploying additional infrastructure?

7.1 Conventional Planning

An engineering team can evaluate configurations manually.

This can work well for small environments, or when the constraints are limited.

As the number of variables and possible configurations grows, however, the number of combinations becomes difficult to evaluate comprehensively.

7.2 Optimization-driven Planning

An optimization engine can represent the relevant infrastructure and workload constraints and evaluate alternative configurations against defined objectives.

The output does not have to be a single opaque answer.

It can instead provide several feasible alternatives:

It is the ability to systematically explore the decision space and expose the trade-offs between alternatives

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

Figure 9: Analysing Different Optimization Options

The value is therefore not simply optimization.

It is the ability to systematically explore the decision space and expose the trade-offs between alternatives

8 Introducing Arctos Labs ECO for Data Centre Optimization

ECO is a generic optimization solution that is built using linear programming, mixed-integer programming, and constraint programming. It is particularly useful when optimization needs to take real-world information into account, requiring data-driven optimizations.

Our approach is particularly attractive where:

- Constraints are well defined
- Decisions are discrete
- Feasibility is important
- Explainability matters

ECO uses a flexible and layered stack where solvers can be replaced and chosen depending on the nature of the optimization problem, including heuristic approaches.

ECO also has a novel framework for processing real-world information and feeding this into the mathematical solver. This is key to make the optimization adapt to new circumstances from surrounding environments as well as to make the processing more efficient.

This framework also gives ECO a unique flexibility in achieving optimization for multiple objectives. ECO's data framework can host multiple optimization models for a variety of optimization objectives, supporting data centre operators to explore the trade-offs.

Arctos Labs has developed several optimization models that are suitable for data centre optimization with the intention to provide such capabilities to DCIM vendors or data centre operators.

8.1 AI and Machine Learning

AI and machine learning can complement optimization by providing predictions.

Examples include:

- Demand forecasting
- Workload forecasting
- Energy consumption prediction
- Thermal prediction
- Failure prediction

The important distinction is:

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152

Prediction and optimization solve different problems

A prediction model can estimate what is likely to happen.

An optimization engine can determine what should be done given that information – decision-making.

Prediction and optimization solve different problems

9 Conclusion

Data centre operations are becoming increasingly complex.

AI workloads, higher IT infrastructure densities, distributed infrastructure, and growing demands on power, cooling, networking, and physical space are making traditional approaches to infrastructure planning more difficult.

At the same time, the resources available to data centre operators remain constrained.

The result is a fundamentally different planning problem.

It is no longer sufficient to understand how much capacity exists.

Operators increasingly need to understand:

Figure 10: The Evolving Planning and Optimization Landscape

DCIM provides an important foundation by creating a structured representation of infrastructure, capacity, and operational conditions. Modern DCIM platforms are already extending into planning and optimization capabilities.

The next step is to connect this infrastructure intelligence with workload information, economic considerations and business objectives.

That is the role of the **data centre optimization layer**.

Its purpose is not simply to monitor infrastructure or automate individual tasks.

Its purpose is to continuously determine:

How should the available infrastructure be configured, allocated, and evolved to achieve the best overall outcome?

How should the available infrastructure be configured, allocated, and evolved to achieve the best overall outcome?

As data centres become more constrained and interconnected, this optimization layer becomes an increasingly important part of the infrastructure stack—bridging the gap between infrastructure visibility, capacity planning and intelligent operations.

Arctos Labs ECO is positioned in that space as it provides a data driven approach that can cater for changing real world conditions, as well as supporting optimization for multiple objectives.

Reach out to us to explore how you can turn your data centre operations more optimal.

Author: Mats Eriksson (assisted by AI) Date: 22 September 2026

Arctos Labs Scandinavia AB, Bangårdsgatan 14, 972 35 Luleå, Sweden. E-mail: info@arctoslabs.com. Web: www.arctoslabs.com. Reg. no: 556743-9152