
aracne

Classificazione Decimale Dewey

— 006.3 (23.) Intelligenza Artificiale e computazione naturale

Thema

— Soggetto: U. Informatica e Tecnologia dell'Informazione
— Qualificatori: 4CT. Per l'istruzione superiore/terziaria/universitaria

GERARDO **IOVANE**
ROLANDO **ROSTOLIS**

DATA CENTERS IN THE AGE OF ARTIFICIAL INTELLIGENCE

COMPUTING INFRASTRUCTURE, DATA ENGINEERING AND ECONOMICS, ENERGY AND ENVIRONMENTAL SUSTAINABILITY AND FUTURE TRENDS

aracne

ISBN
979-12-218-2707-1

PRIMA EDIZIONE
ROMA 4 MAGGIO 2026

Table of Contents

Foreword

This book stems from a dual need: scientific and educational on the one hand, and practical on the other. The scientific need is for a text that addresses, with analytical rigor, engineering depth, and systematic breadth, the technological infrastructure that today underpins the entire global digital economy—the modern data center, in its most advanced and demanding form, as dictated by the workloads of artificial intelligence. The educational need is to offer doctoral students, researchers, and industry professionals a tool capable of navigating with equal competence between the mathematical formalization of inter-layer relationships, the engineering specification of cooling architectures, the economic analysis of financing models, and the geopolitical framework of digital sovereignty, so as to produce professionals ready for implementation.

Anyone who has worked with data, machine learning models, distributed systems, or large-scale networks knows the dizzying sensation one feels when one truly understands—not merely intuitively, but quantitatively and structurally—how many physical resources, how much energy, and how much engineering complexity lie behind a single training cycle of a Large Language Model or behind the uninterrupted flow of inference requests that supports a cloud service on a planetary scale. A model with 700 billion parameters trained on tens of thousands of GPUs for months consumes energy in the order of gigawatt-hours; a hyperscale data center at full capacity can draw hundreds of megawatts from the power grid; the latency of a single packet between servers in different racks, when multiplied by the millions of exchanges occurring every second in a GPU cluster, becomes the bottleneck that determines the efficiency of the entire training system. These realities are not marginal: they are central, structural, and deserve treatment commensurate with their complexity.

The ideal reader of this book is diverse. It could be a graduate student in computer science, computer engineering, electrical engineering, or civil engineering who needs a systemic view of the physical infrastructure that hosts their computational experiments. It could be a researcher in distributed systems, cloud computing, or artificial intelligence who wishes to understand the formal relationships between resource constraints at the lower levels of the technology stack and the observable

performance at the application- s levels. It could be the professional who designs, manages, or finances data centers and seeks a reference that integrates engineering, computer science, and economic literature. It could be the graduate student in STEM disciplines who wants to build a deep and scalable understanding of digital infrastructure before specializing in one of its constituent dimensions. In each of these cases, the book offers differentiated levels of reading: mathematical formalism is always present for those who seek it, but the text is structured so that conceptual understanding is accessible even to those approaching certain engineering aspects for the first time.

A word on the method. This book was conceived as an advanced training tool that measures up to empirical evidence, peer-reviewed scientific literature, and the technical specifications of industry operators. The bibliographic citations are numerous and precise; the formulas are derived from principles and explained step by step; the operational data are drawn from primary and institutional sources of recognized authority. We have chosen not to sacrifice analytical depth on the altar of immediate accessibility, convinced that premature simplification is the greatest obstacle to the development of robust and transferable skills. At the same time, the thematic progression has been constructed with pedagogical care, so that each chapter consciously builds upon the foundations established in the preceding ones and prepares the conceptual groundwork for the subsequent ones.

Data centers are not merely machines. They are the result of engineering, economic, political, and environmental choices layered over time. Understanding how they work—truly, in their formal detail—is today a fundamental skill not only for those who design or manage them, but for anyone who wishes to understand the structural transformations that the digital economy and artificial intelligence are imposing on global society. This volume is a contribution to that understanding.

Introduction

This book is structured as a dual architecture: one formal and one systemic.

The formal architecture is established from the outset: the book introduces, capitalizes on, and applies a coherent set of mathematical and engineering tools—energy consumption models, inter-layer constraint functions, thermal and computational efficiency metrics, and frameworks for economic and risk analysis—which are progressively developed and reused throughout the forty-four chapters that make up the work. This formal architecture is not ornamental: it is the skeleton that ensures the analytical coherence of the entire text and allows for a seamless transition from the thermodynamics of cooling systems to the modeling of GPU clusters, from the analysis of switch fabrics to the evaluation of the cost of capital in investment models.

Systemic architecture is the choice to treat the data center not as an aggregate of independent components but as a complex cyber-physical system in which every level—from the civil structure and seismic foundations to the software layers for container orchestration and the data pipelines for machine learning—is formally connected to the others through relationships of dependency, constraint, and optimization that must be understood and quantified in order to design, manage, and scale the infrastructure efficiently and sustainably. The central theoretical contribution of this volume is a six-level framework—the AI Data Center Architecture Stack—equipped with five formal constraint functions: $\phi_{5\to6}$,$\phi_{4\to5}$,$\phi_{3\to4}$,$\phi_{2\to3}$,$\phi_{1\to2}$. These specify the minimum dependency relationships between adjacent levels and allow the minimum provisioning requirements for each layer to be propagated downward—from the AI workload specified at level 6 down to the physical infrastructure at level 1. This framework, developed based on industry literature and empirically validated against publicly disclosed operational data from five Western hyperscale operators, serves as the analytical backbone of the entire work.

Structure and Organization of the Volume

The volume is organized into thirteen thematic parts comprising a total of forty-four chapters, plus a global cartographic appendix. This structure reflects the intrinsic complexity of the subject matter: a modern data center integrates expertise in computer science, civil and structural engineering, electrical and power systems