
aracne

Classificazione Decimale Dewey

— 004.6 (23.) INTERFACCIAMENTO E COMUNICAZIONI

Thema

— Soggetto: U. Informatica e Tecnologia dell'Informazione
— Qualificatori: 3MR. XXI secolo, 2000-2100

GERARDO **IOVANE**

EUROPEAN AND NATIONAL DATA CENTERS

DISTRIBUTED, MULTI-SCALE, CONFEDERATE HYPERSCALE FACILITIES

aracne



©

ISBN
979-12-218-3089-7

PRIMA EDIZIONE
ROMA 16 NOVEMBRE 2026

Table of Contents

Preface 9

Introduction 13

CHAPTER 1 — THE DATA CENTER AS A FUNDAMENTAL INFRASTRUCTURE OF DIGITAL CIVILIZATION: DATA, COMPUTING, ARTIFICIAL INTELLIGENCE, AND THE LIMITS OF CONCENTRATION 25

CHAPTER 2 — HISTORY OF CENTRALIZATION AND DECENTRALIZATION IN COMPUTING: FROM THE MAINFRAME TO THE MULTISCALE CONFEDERATION 43

CHAPTER 3 — WHY HYPERSCALE DOES NOT HAVE TO CORRESPOND TO A BUILDING: PHYSICAL SCALE, LOGICAL SCALE, AND EMERGENT PROPERTIES 59

CHAPTER 4 — THE COOPERATIVE DIGITAL INFRASTRUCTURE AS THE THEORETICAL FOUNDATION OF THE CONFEDERATION 75

CHAPTER 5 — FORMALIZATION OF THE MULTISCALE CONFEDERATED HYPERSCALE: STRUCTURE, RESOURCE VECTORS, AND DERIVED CAPACITIES 91

CHAPTER 6 — C-ADCAS: THE CONFEDERATED AI DATA CENTER ARCHITECTURE STACK 107

CHAPTER 7 — COOPERATIVE SCALING LAWS: CAPACITY, AVAILABILITY, DIVERSITY, AND SATURATION OF COOPERATION 123

CHAPTER 8 — COMPUTATIONAL COUPLING THEORY: THE κW INDEX AND WORKLOAD CLASSIFICATION 139

CHAPTER 9 — EFFECTIVE CONFEDERATED COMPUTE: FROM INSTALLED COMPUTING POWER TO ACTUALLY COOPERATIVE COMPUTING 157

CHAPTER 10 — THE 100-kW MICRO DATA CENTER: ARCHITECTURE, OPERATIONAL AUTONOMY, AND THE ECONOMICS OF PROXIMITY 173

CHAPTER 11 — THE APPROXIMATELY 1 MW MINI DATA CENTER: TERRITORIAL AGGREGATION, INDUSTRIAL ARTIFICIAL INTELLIGENCE, AND ENERGY COMMUNITIES 189

CHAPTER 12 — THE TENS-OF-MEGAWATTS MEDIUM-SIZED DATA CENTER: REGIONAL CLOUD, METROPOLITAN MODELING FACILITY, AND APPLIED SOVEREIGNTY 205

CHAPTER 13 — THE LARGE DATA CENTER UP TO APPROXIMATELY ONE GIGAWATT: CAMPUSES, MEGACLUSTERS, AND THE CONFEDERATION'S COMPUTING HUB 221

CHAPTER 14 — OPTIMAL HYBRID ARCHITECTURES: SYSTEMATIC COMPARISON AT EQUAL INSTALLED CAPACITY 237

CHAPTER 15 — ITALIAN NATIONAL ARCHITECTURE WITH 6,012 NODES: TOPOLOGY, LOGICAL HIERARCHY, AND LOAD ROUTING 255

CHAPTER 16 — COMPUTATIONAL CAPACITY OF THE ITALIAN SCENARIO: AGGREGATE POWER, SENSITIVITY ANALYSIS, AND ACTUALLY COOPERATIVE CALCULATION 271

CHAPTER 17 — NATIONAL CONFEDERATION VS. TRADITIONAL HYPERSCALE: WHEN THE MONOLITHIC ARCHITECTURE WINS, WHEN THE FEDERATED ARCHITECTURE WINS, AND WHEN THE HYBRID ARCHITECTURE WINS 287

CHAPTER 18 — NATIONAL AI COMPUTE BACKBONE: NATIONAL COMPUTING BACKBONE, MODEL DISTRIBUTION, AND SYSTEM INFERENCE 305

CHAPTER 19 — FROM SITE SELECTION TO NETWORK-OF-SITES SELECTION: LOCATION AS A NETWORK PROBLEM 321

CHAPTER 20 — LOCATION MATHEMATICS: FACILITY LOCATION MODELS, GRAPH PARTITIONING, AND MULTI-OBJECTIVE OPTIMIZATION 337

CHAPTER 21 — INLAND AREAS AS PRODUCTIVE DIGITAL INFRASTRUCTURE: COMPUTING, EMPLOYMENT, AND TERRITORIAL COHESION 353

CHAPTER 22 — THE POWER OF CONFEDERATED HYPERSCALE: LOAD PROFILE, DIVERSITY, AND NETWORK SERVICES 371

CHAPTER 23 — DISTRIBUTED RENEWABLE ENERGY SOURCES: PHOTOVOLTAIC, WIND, HYDROELECTRIC, AND PORTFOLIO OPTIMIZATION 389

CHAPTER 24 — ELECTROCHEMICAL STORAGE SYSTEMS: FROM UNINTERRUPTIBLE POWER SUPPLIES TO FEDERATED STORAGE 407

CHAPTER 25 — ENERGY COMMUNITIES AND DATA COMMUNITIES: THE DIGITAL PROSUMER AND CONTRIBUTION ACCOUNTING 425

CHAPTER 26 — CARBON-AWARE LOAD SHIFTING: TEMPORAL AND GEOGRAPHICAL SCHEDULING AND LIMITATIONS OF THE METHOD 443

CHAPTER 27 — THERMODYNAMICS OF DISTRIBUTED COMPUTING: DENSITY, COOLING TECHNOLOGIES, AND SCALING CONSTRAINTS 461

CHAPTER 28 — HEAT RECOVERY AND DISTRICT HEATING: THERMAL QUALITY, SCALE OF CONSUMERS, AND CONTRACTUAL MODELS 477

CHAPTER 29 — CLIMATE-AWARE COOLING AND THERMAL ORCHESTRATION: METRICS, TRADE-OFFS, AND LIMITATIONS 493

CHAPTER 30 — THE NETWORK AS A COMPUTATIONAL COMPONENT: TOPOLOGY, SCALING, AND STRUCTURAL COST OF DISTRIBUTION 511

CHAPTER 31 — TRANSPORT TECHNOLOGIES AND CONFEDERAL FABRIC: OPTICS, DETERMINISTIC LATENCY, AND SEGMENTATION 527

CHAPTER 32 — JOINT ORCHESTRATION OF NETWORK AND COMPUTING: CO-SCHEDULING, RESERVATION, AND CONTROLLED DEGRADATION 543

CHAPTER 33 — CONFEDERATED COMPUTING: HETEROGENEITY, RESOURCE ABSTRACTION, AND SHARED POOLS 559

CHAPTER 34 — FEDERATED STORAGE: REPLICATION, ENCODING, CONSISTENCY, AND DISTRIBUTED DURABILITY 575

CHAPTER 35 — SOVEREIGN DATA: CLASSIFICATION, ELIGIBILITY CRITERIA, AND THE PRICE OF SOVEREIGNTY 591

CHAPTER 36 — ARTIFICIAL INTELLIGENCE IN CONFEDERATED HYPERSCALE: MODEL LIFECYCLE AND DISTRIBUTION ACROSS SCALES 609

CHAPTER 37 — ARTIFICIAL INTELLIGENCE AS A TOOL FOR INFRASTRUCTURE GOVERNANCE: FORECASTING, ALLOCATION, AND UNCERTAIN DECISION-MAKING 625

CHAPTER 38 — AGENTS AND MULTI-AGENT SYSTEMS FOR CONFEDERAL ORCHESTRATION: NEGOTIATION, TRUST, AND ACCOUNTABILITY 641

CHAPTER 39 — SOVEREIGN ARTIFICIAL INTELLIGENCE AND TERRITORIAL MODELS: CHAIN OF DEPENDENCE, STRATEGIES FOR AUTONOMY, AND PLURALISM 659

CHAPTER 40 — THE CONFEDERATION'S OPERATING SYSTEM: CONTROL PLAN, IDENTITY, AND INTEROPERABILITY 677

CHAPTER 41 — MULTI-OBJECTIVE SCHEDULING AND ALLOCATION: SERVICE CLASSES, FEASIBLE HEURISTICS, AND FAIRNESS 693

CHAPTER 42 — INTEROPERABILITY, STANDARDS, AND CONFEDERATION INTERFACES: VOCABULARIES, COMPLIANCE, AND GOVERNANCE 709

CHAPTER 43 — SECURITY OF THE CONFEDERAL INFRASTRUCTURE: THREAT MODEL, ZERO TRUST, AND HOSTILE PARTICIPANT 727

CHAPTER 44 — RESILIENCE AND CONTINUITY: CORRELATED FAILURES, COOPERATIVE REDUNDANCY, AND CONTROLLED DEGRADATION 743

CHAPTER 45 — EXPERIMENTAL VERIFICATION OF RESILIENCE: CHAOS ENGINEERING, IMPACT RADIUS, AND EXPERIMENT MANAGEMENT 759

CHAPTER 46 — INCIDENT MANAGEMENT AND ORGANIZATIONAL LEARNING: DIAGNOSIS, COMMAND, AND SYSTEMIC FACTORS 775

CHAPTER 47 — CONFIDENTIAL COMPUTATION AND IN-USE DATA PROTECTION: TRUSTED ENVIRONMENTS, ADVANCED CRYPTOGRAPHY, AND TECHNIQUE SELECTION 791

CHAPTER 48 — QUANTUM RESILIENCE: THREATS, POST-QUANTUM MIGRATION, AND KEY ARCHITECTURE 807

CHAPTER 49 — DEMAND GROWTH AND MATERIAL CONSTRAINTS: REBOUND, SUFFICIENCY, AND CAPACITY ALLOCATION 823

CHAPTER 50 — PLANNING AND INVESTMENT: CAPACITY DESIGN, ECONOMIC EVALUATION, AND IMPLEMENTATION SEQUENCE 839

CHAPTER 51 — GOVERNANCE OF THE CONFEDERATION: ACTORS, RULES, AND DECISION-MAKING PROCESSES 855

CHAPTER 52 — ECONOMICS OF THE FEDERATION: RESOURCE CREDITS, TARIFFS, AND CONDITIONS OF PERSISTENCE 871

CHAPTER 53 — THE REGULATORY FRAMEWORK: APPLICABLE REGULATIONS, GAPS, AND NECESSARY ADAPTATIONS 887

CHAPTER 54 — THE INSTITUTIONAL FRAMEWORK: COMPETENCIES, MULTILEVEL COORDINATION, AND A VIABLE PATH 903

CHAPTER 55 — THE EUROPEAN DIMENSION: WHAT TO FEDERATE ON A CONTINENTAL SCALE AND UNDER WHAT GOVERNANCE 919

CHAPTER 56 — SCENARIOS: TRAJECTORIES, FORKS, AND ROAD MAPS 935

CHAPTER 57 — THE TERRITORIAL DIGITAL PROSUMER: CONTRIBUTION, CREDITS, AND LOCAL VALUE CAPTURE 951

CHAPTER 58 — THE CONFEDERATION AS INDUSTRIAL POLICY: SOVEREIGNTY, COHESION, AND STRATEGIC AUTONOMY 969

CHAPTER 59 — EUROPEAN ARCHITECTURE OF NATIONAL CONFEDERATIONS: INTEROPERABILITY WITHOUT LOSS OF SOVEREIGNTY 987

CHAPTER 60 — ERASMUS: THE EUROPEAN COMPUTING CONFEDERATION 1007

CHAPTER 61 — GLOBAL BENCHMARK: A COMPARISON OF TRADITIONAL HYPERSCALE PROVIDERS, NATIONAL CONFEDERATIONS, AND ERASMUS 1025

CHAPTER 62 — THE CONFEDERATED HYPERSCALE SCORE: *CHSCompute*, *CHSSystem* , *CHSAI*, AND OVERALL SCORE 1043

CHAPTER 63 — THE LARGE-SCALE COMPARATIVE EXPERIMENT: SIX SCENARIOS WITH EQUAL CAPACITY AND THIRTY-THREE METRICS 1059

CHAPTER 64 — THE CONFEDERATED HYPERSCALE DOMINANCE MAP: WHEN EACH SCALE DOMINATES 1077

CHAPTER 65 — THE OPTIMAL SCALE THEOREM: STATEMENT, VALIDITY CONDITIONS, AND DESIGN IMPLICATIONS 1093

CONCLUSIONS 1111

APPENDIX A — FORMAL MATHEMATICAL MODEL OF THE CONFEDERATION 1119

APPENDIX B — EMERGENT RELIABILITY AND AVAILABILITY 1125

APPENDIX C — MODELS OF COMPUTATIONAL COUPLING 1131

APPENDIX D — MODELS OF ACTUAL COOPERATIVE CALCULATION 1135

APPENDIX E — ENERGY MODELS 1139

APPENDIX F — THERMAL MODELS 1143

APPENDIX G — NETWORK MODELS 1147

APPENDIX H — ARTIFICIAL INTELLIGENCE LOAD MODELS 1151

APPENDIX I — CAPITAL, OPERATING, AND TOTAL COST MODELS 1157

APPENDIX J — THE APPLICABLE REGULATORY FRAMEWORK 1161

APPENDIX K — SIMULATION METHODOLOGY AND REPRODUCIBILITY 1167

APPENDIX L — REFERENCE ARCHITECTURES FROM 100 kW TO 1 GW 1173

APPENDIX M — ITALIAN ATLAS: STRUCTURE, VARIABLES, AND SOURCES 1179

APPENDIX N — EUROPEAN ATLAS: STRUCTURE, VARIABLES, AND SOURCES 1185

Preface

The volume preceding this one, Data Centers in the Age of Artificial Intelligence, concluded with an observation that at the time seemed merely a limitation of the method but which, over time, has revealed itself to be the beginning of a new problem. After describing the data center in its physical form—from site selection to the structural framework, from electrical distribution to the cold chain, from accelerator clusters to interconnection networks—and after discussing its economics, sustainability, and regulation, one question remained that the framework of that work could not address: to what extent does it make sense to continue studying the data center as an object, when the factors that determine its behavior are now systemic in nature? Available power is not a property of the facility itself but of the electrical grid node to which the facility is connected. The latency perceived by the user does not depend on the site's internal architecture but on its location relative to demand. Sustainability is not limited to the ratio of energy consumed to useful energy, because water consumption, land use, heat dissipation, and embodied carbon are quantities measured across the entire territory and not just within the perimeter of the fence. Finally, service availability has long been an emergent property of sets of sites, not of the individual site, and no serious designer calculates it based solely on local redundancy anymore.

This volume arises from these observations. Its working hypothesis is simple to state but challenging to defend: the relevant unit of analysis for contemporary computational infrastructure is no longer the data center, but the confederation of data centers of varying scales that cooperate across a territory. This is not merely a descriptive shift in perspective. If the unit of analysis changes, the metrics used to measure it change, the design criteria change, the cost models change, and the way a nation-state conceives of its computing capacity changes. A volume dedicated to this shift is therefore not an appendix to the previous one, but its necessary evolution, when one seeks a vision that harmonizes with the territory.

The immediate impetus came from research conducted in recent years on cooperative digital infrastructures, federated micro data centers, and the mechanisms that enable their distributed governance under conditions of limited trust. That work, which originated in a seemingly tangential context, has shown that the theoretical problems raised by the federation of small nodes are, in essence, no different from those raised by the federation of large nodes: it is always a matter of determining when the sum of capacities is operationally significant, at what cost cooperation can take place, and what properties emerge from the whole that none of the individual parts possesses. Extending those questions to the national and continental scales was the next step, and this book is the result.

It is important to state right away what position this work takes in the public debate on the concentration of computing, because that debate is highly polarized today, and polarization leads to poor engineering. On the one hand, it is argued that only mega-facilities allow for the training of models that define the technological frontier, and that any discussion of

distribution is merely a form of consolation for those who lack the necessary capital. On the other hand, it is argued that centralization is a mistake in and of itself—environmentally unsustainable, socially unacceptable, and technically surmountable by a sufficiently dense network of small nodes. Neither position stands up to quantitative scrutiny. The first ignores that a very large portion of actual computational demand is weakly coupled, latency-sensitive, and constrained by data locality—and that for that portion, mega-facilities are a costly and inefficient solution. The second ignores that there are classes of workloads whose execution requires physical proximity between accelerators that no geographic network, however capable, can substitute for. The thesis of this book lies elsewhere and is more demanding than both: every infrastructure scale has its own domain of optimality; that domain can be determined based on the characteristics of the workload and the context; and the task of infrastructure engineering is not to choose a single scale but to orchestrate them all within a single architecture.

This position has a consequence worth anticipating. The book does not pit "confederation" against "hyperscale": confederation, as defined here, encompasses hyperscale as its highest level. Anyone looking for an argument against large computing campuses in these pages will be disappointed, just as those seeking proof that territorial capillarity is a value in and of itself will be disappointed. The book's first key thesis is that the term "hyperscale" can describe a distributed system and not just a building; the second is that geography, energy, the network, and the territory should be treated as components of computational architecture, not as external constraints imposed on the design.

The book is aimed at four types of readers. Researchers will find a formal framework that quantifies concepts previously addressed in a discursive manner, particularly those of computational coupling, effective cooperative capacity, and sufficient scale. Engineers and designers will find sizing criteria, systematic comparisons between node classes, and an explicit discussion of the points at which geographic distribution ceases to be advantageous. Economists and public policymakers will find an analysis of costs, ownership models, and industrial policy implications, along with an overview of the conditions under which computing capacity becomes an attribute of sovereignty and not merely a budget line item. Finally, local administrators will find the rationale for why a computing node installed in a marginal region can generate measurable value, along with the conditions necessary for this to actually happen rather than remain merely a promise.

A methodological caveat is in order right from the start. The book makes extensive use of national and European numerical scenarios, designed to allow for the comparison of alternative architectures with the same installed capacity. These scenarios are not forecasts, they are not investment plans, and they are not intended to suggest that a particular configuration must be implemented. They are tools for reasoning and should be read as such: their function is to show how different variables behave when the same capacity is organized in different ways, not to indicate how many facilities a country should build. Anyone who

were to extract figures from these pages to use as planning targets would be misusing them—a practice against which the text repeatedly warns.

A debt of gratitude remains. The work that led to this volume was nourished by daily exchanges with colleagues and collaborators on the team coordinated by the author, who agreed to discuss its hypotheses when they were still fragile and who pointed out its weaknesses with the candor that only mutual respect makes possible. Special thanks go to the university environment in which the research took place, which guaranteed the freedom to pursue a path not yet well-established, and to the students in the digital infrastructure courses, whose insistence on asking why every threshold and every coefficient was set as it was improved the formulation of many of the ideas found here. Thanks also go to the technicians, designers, and operators who agreed to describe real-world operational problems: contact with practical reality prevents theory from becoming merely elegant and useless. Any errors, oversimplifications, and questionable positions that the reader may find in these pages remain, of course, the sole responsibility of the author.

One final observation concerns the timing of this book's publication. Computing infrastructure is undergoing a phase of expansion that is unprecedented in terms of capital intensity and the speed of obsolescence; any text that purports to describe it runs the risk of becoming outdated before it even reaches the reader. To mitigate that risk, this volume deliberately prioritizes structural parameters over short-term data, relationships between quantities over absolute values, and decision criteria over the technological solutions of the moment. Generations of accelerators will change, costs per unit of capacity will fall, and power densities per rack will continue to rise; what will not change—at least as long as the speed of light remains what it is—is the fact that communication takes time and that time spent communicating is not time spent computing. The entire argumentative framework of this volume rests on this fundamental asymmetry.

Artificial intelligence tools were used in the creation of this work to assist with translation, document research, and stylistic revision of the text. The author personally verified, reviewed the content, and provided final approval of the text.

Introduction

I.1 The Context: The Structural Transformation of the Global Digital Infrastructure

The global digital infrastructure is undergoing its most profound transformation since the advent of the cloud, and the nature of this transformation is often described inaccurately. It is not merely a matter of quantitative growth in installed capacity, however impressive and unprecedented that growth may be in the history of technical infrastructure. It is a qualitative shift in demand, which simultaneously alters the power density of systems, the cost structure, location constraints, and the relationship between digital infrastructure and the power grid. Each of these changes, taken in isolation, would be manageable using the established tools of the field. It is their simultaneity that creates the discontinuity.

The first element of the transformation is the changed composition of the load. For over a decade, the dominant demand has been for general-purpose cloud services, characterized by moderate computational intensity, relatively high tolerance for latency, and predictable temporal distribution. The industrial-scale adoption of machine learning workloads has introduced a different kind of demand—time-concentrated, extremely intensive in its use of accelerators, and characterized by internal communication requirements that have no equivalent in traditional computing. The immediate consequence was an increase in power density per cabinet, which grew by an order of magnitude over the course of just a few years, with cascading effects on the electrical architecture, the cold chain, and the structural design of data centers.

The second factor is the emergence of energy constraints as the primary constraint. For a long time, the availability of electrical power was just one design parameter among others, easily resolved with an adequate utility connection. Today, the capacity to connect to the transmission grid has become the determining factor in where a facility can be built and when it can become operational, with waiting times in various European regions exceeding the duration of the facility's construction itself. This shift has implications that go beyond logistics: it means that digital infrastructure is now in direct competition with other uses of electricity, and that its development now depends on planning decisions that lie outside the digital sector.

The third factor is the emergence of environmental and land-use constraints as binding conditions rather than merely ancillary requirements. Water consumption by cooling systems, land use, impact on the landscape, contribution to climate-changing emissions, and disputes with local communities over permitting processes are no longer marginal issues in the implementation phase; they have become factors that determine the very feasibility of large-scale projects. In areas with the highest concentration of infrastructure, social resistance to the establishment of new campuses has reached such intensity that it has led to administrative moratoriums, with direct effects on the geography of investments.

The fourth element is the transformation of computing capacity into an attribute of sovereignty. When computing becomes the productive factor underpinning scientific research, defense, healthcare, public administration, and a growing share of economic activity, its geographic and legal location ceases to be irrelevant. Data localization policies, supplier certification requirements, and restrictions on the export of advanced components have led to a fragmentation of the global market that no operator can ignore. The computational capacity available within a jurisdiction and under that jurisdiction's legal control has become a matter of public interest.

The fifth element—less discussed but no less significant—concerns the structure of regional demand. The growth of digital services is not distributed uniformly, nor is it concentrated where the infrastructure is concentrated. Applications that require short response times and service continuity arise where people live and businesses operate—that is, widely across the territory—while infrastructure is concentrated in a limited number of hubs chosen for reasons of energy costs, connectivity availability, and tax regimes. This disconnect between the geography of demand and the geography of supply results in costs, which manifest as latency, transit traffic, dependence on transport infrastructure, and systemic fragility.

Together, these five elements define the context within which the subject of this volume is situated. Computational demand is growing faster than the capacity to absorb it under the concentration paradigm; the physical constraints limiting concentration cannot be overcome by technological innovation in the short term; real demand is geographically distributed; and the institutional dimension of computing capacity imposes location criteria that purely economic optimization does not take into account. Under these conditions, the question is not whether the infrastructure will become distributed—because distribution is already underway, driven by these constraints—but whether this distribution will occur in a governed and planned manner or in an incremental and inconsistent one. This volume aims to provide the conceptual tools for the former outcome.

I.2 From the Data Center to the Confederation: Why the Mega-Facility Paradigm Does Not Fully Address the Problem

The dominant paradigm of the past fifteen years has equated the progress of computational infrastructure with the increase in the size of individual facilities. This identification was not arbitrary: it reflects real and documented economies of scale concerning the unit cost of energy, the efficiency of cooling systems, the productivity of technical staff, bargaining power in the procurement of components, and the ability to co-design hardware and software at scales that justify the development of dedicated silicon. Those who argue that this concentration was a mistake ignore the robustness of these economies of scale. The point is not to deny them, but to recognize that every scale-return curve has a terminal point where marginal benefits dry up while marginal costs continue to rise—and that digital infrastructure has reached that point.

The signs of this exhaustion are recognizable. The marginal cost of electrical connection grows more than proportionally with scale, because beyond a certain threshold, no network node is capable of absorbing the load without reinforcement measures that entail their own time and cost. The marginal cost of cooling increases with density, because the transition from air to liquid cooling entails a discontinuity in both plant design and organizational structure. The marginal cost of local consent increases with concentration, because the impact perceived by a community is not a linear function of installed capacity. The marginal cost of risk increases with capacity concentration, because the loss of a site hosting a significant share of national capacity causes systemic damage that cannot be compensated for. None of these effects is new; what is new is that they are all occurring simultaneously and within the same size range.

This observation is accompanied by an argument of a different nature, one that concerns not costs but suitability. A portion of computational demand is inherently non-relocatable: clinical data that cannot leave the healthcare facility; industrial processes requiring response times on the order of milliseconds; urban surveillance systems that generate data streams whose full transmission to a remote site would be more expensive than their local processing; and public services that must continue to function when geographic connectivity is interrupted. For this portion of the demand, a remote data center is not just a costly solution—it is an impractical one. The distributed infrastructure is therefore not a choice, but the necessary response to a specific set of requirements.

The edge computing paradigm has recognized this need but has addressed it only partially. The edge was conceived as a peripheral extension of the cloud—a thin, proximity-based layer that filters, aggregates, and pre-processes data before sending it to the center. This conception preserves the hierarchy intact: the center computes, the periphery prepares. It does not address the question at hand, which concerns the possibility that nodes of different scales might cooperate on a plane of substantial functional parity, each contributing to a single system capacity rather than occupying fixed positions in a pyramid.

The notion of a confederation arises from this divergence. A confederation of data centers is not a hierarchy with a vertex and a base, but a network in which the workload can move upward toward larger-scale nodes, downward toward smaller-scale nodes, or laterally between nodes of equivalent scale, depending on the requirements of the workload itself and the state of the system. The large-scale node remains indispensable for what only it can do, but it ceases to be the logical center of the architecture and instead becomes one of its layers. The capacity offered by the system is neither the capacity of the largest node nor the indiscriminate sum of the nodes' capacities, but a quantity that depends on how effectively the nodes can cooperate on a given load.

This shift gives rise to the need for a new conceptual framework, because established metrics are unable to describe the system. The energy efficiency of the infrastructure—measured as the ratio of total energy consumed to useful energy—is defined within the boundaries of a single facility and loses its meaning when those boundaries are distributed across a territory.

Service availability, calculated based on the redundancy of individual facilities, does not capture the redundancy that arises from the multiplicity of sites. Computing capacity, expressed as the number of operations per second installed, says nothing about the fraction of those operations that can contribute to the same task. We need metrics that measure cooperation, not installation, and this is why this volume devotes its second part to developing such a framework before addressing any engineering issues.

I.3 The book's central thesis, its corollaries, and its stated limitations

The central thesis of this volume can be stated in a single proposition: hyperscale is not necessarily a dimensional property of the individual data center, but can constitute an emergent property of the confederation. The adjective that for a decade has described buildings and campuses now, in this perspective, describes a systemic behavior, recognizable when a set of nodes cooperates in such a way as to offer capacity, availability, and flexibility that none of the nodes possesses individually. This statement is not metaphorical. It is verifiable, and the book identifies the conditions under which it is true and those under which it is false.

The complementary thesis concerns the status of the physical and institutional context. Geography, energy, the network, and the territory are not external constraints imposed on the design, but components of computational architecture, on par with memory and the interconnection fabric. This assertion has precise operational consequences: the distance between two nodes is an architectural parameter in the same sense that memory access latency is; the hourly profile of energy prices is a scheduling parameter in the same sense that process priority is; the jurisdiction within which data resides is an allocation constraint in the same sense that a node's remaining capacity is. Treating these elements as context rather than as components produces architectures that work on paper but fail in practice.

A strategic principle emerges from this thesis that deserves to be emphasized to avoid recurring misunderstandings: the distributed paradigm does not eliminate large-scale data centers; it incorporates them. The confederation includes nodes with capacities of hundreds of megawatts as its constituent elements and recognizes their irreplaceability for the class of workloads that require physical proximity among a large number of accelerators. Anyone who reads the book as a manifesto against centralization misunderstands its argument. The position taken is that centralization is necessary for one part of the demand and inefficient for the remaining part; the task of the architecture is to recognize the boundary between the two.

The operating principle underpinning the entire framework is that of the "minimum sufficient scale": each workload must be executed at the smallest infrastructural scale capable of simultaneously meeting its requirements for computation, memory, bandwidth, latency, resilience, sustainability, and sovereignty. The principle is normative, not descriptive: it indicates how the load should be allocated, not how it is allocated in current practice. Its application requires that the requirements be expressible in quantitative terms

and that the capabilities of the node classes be known—conditions that this volume seeks to make verifiable.

There are three corollaries to this thesis. The first is that there is no universally optimal scale, but rather an optimal scale that depends on the workload and context, which can be determined through an explicit procedure. The second is that the benefit of federation is not uniform across workloads, but increases as coupling decreases and as sensitivity to latency, data locality constraints, and the weight of energy and geographical factors increases. The third is that cooperation has a cost, which manifests itself in orchestration complexity, geographic traffic, and governance overhead; this cost defines a threshold beyond which adding nodes to the federation no longer produces value.

Finally, the limits of validity must be stated. This volume presents arguments on a theoretical and modeling basis, not an experimental one: the architectures compared are analytical constructs evaluated through models, not actual systems that have been built and measured. The model parameters are derived from technical literature, public documentation from operators, and design experience, with a degree of uncertainty that is considerable in some cases and is disclosed on a case-by-case basis. The conclusions regarding costs depend on assumptions about energy, capital, and hardware prices that are valid at the time of writing but may not hold true beyond that point. The conclusions regarding sustainability depend on the quality of available data on embedded emissions, which remains unsatisfactory. Finally, the book adopts a European perspective; its institutional and regulatory conclusions are not transferable to other legal systems without adaptation. Stating these limitations does not weaken the argument: it defines its scope of applicability, which is what distinguishes a scientific hypothesis from an ideological position.

I.4 The Fundamental Quantities of Distributed Hyperscale: Coupling, Cooperative Computation, Sufficient Scale, and Confederation Score

The theoretical framework of this volume rests on a small number of quantities, introduced here in their essential form and systematically discussed in Part Two. The first and most important is the computational coupling index, denoted by$\kappa_W \in [0,1]$ and referred to a workload *W*. It measures the degree to which the execution of the workload requires communication and synchronization among the participating computing units. A value close to one characterizes strongly coupled workloads, where each processing step is preceded and followed by an exchange of information among all involved units , as occurs in the training of frontier models and in fine-grained scientific simulations. A value close to zero characterizes workloads in which units operate largely independently, such as inference on separate requests, content distribution, or the processing of sensor streams. The index is not a qualitative label: it is estimated based on observable quantities, foremost among which is the ratio of time spent on communication to time spent on computation.

This coupling gives rise to the criterion governing geographical distributability. Since the latency between distant nodes exceeds the latency within a single node by three orders of

magnitude, a workload with highκ_W suffers a penalty in distribution that can entirely negate the advantage of additional capacity, whereas a workload with lowκ_W suffers only a negligible penalty. This asymmetry, which stems from a physical property rather than an insurmountable technological limitation, is the foundation of this entire volume and the reason why the confederation cannot be presented as a universal solution.

The second quantity quantifies the distinction between installed capacity and usable capacity. The actual cooperative calculation is expressed as

$$\mathrm{ECC} = C_{\mathrm{raw}} \cdot \eta_{\mathrm{HW}} \cdot \eta_{\mathrm{util}} \cdot \eta_{\mathrm{net}} \cdot \eta_{\mathrm{place}},$$

whereC_{raw} is the theoretical installed capacity,η_{HW} is the hardware's efficiency relative to its peak value,η_{util} is the actual utilization,η_{net} is the factor summarizing the penalty due to the network, andη_{place} is the quality of the allocation performed by the orchestrator. The multiplicative structure is significant: since each factor is less than one, a system may have an impressive nominal capacity and a modest cooperative capacity, and the gap between the two grows as the load is coupled. Alongside this metric is the distributed service capacity,$Q_{\mathrm{tot}} = \sum_i Q_i$, which instead measures the overall capacity to serve independent loads and for which aggregation is legitimate.

The third metric formalizes the operational principle of volume. The sufficient scale for a load is defined as

$$s^{(W)} = \min_s, s: C_s \geq C_W, ; M_s \geq M_W, ; B_s \geq B_W, ; L_s \leq L_W, ; R_s \geq R_W, ; S_s \geq S_W,$$

where the symbols denote, respectively, the requirements for computation, memory, bandwidth, latency, resilience, and sovereignty, compared to the corresponding capabilities of node class s. The formulation is deliberately elementary, because its difficulty lies not in the mathematics but in determining the requirements and capabilities. In a more general form, the optimal scale is written as

$$\mathrm{Scale}^*(W) = f(\kappa_W, L_W, B_W, C_W, M_W, D_W, E, S_W, R_W, \mathrm{Cost}),$$

an expression that captures the dependence on the workload and the context.

The fourth parameter concerns availability. For nodes with independent failures, system availability is written as

$$A_{\mathrm{sys}} = 1 - \prod_i (1 - A_i),$$

an expression that grows rapidly with the number of nodes and explains why the multiplicity of sites effectively replaces facility redundancy. The assumption of independence, however, is the weakest link in the entire framework, and the book devotes a specific analysis to it: when failures are correlated—due to a shared power grid, connectivity, or exposure to

natural events—the advantage of distribution is reduced to an extent that must be estimated rather than assumed.

The final metrics are summary measures. The "CFSE " metric measures the sustainability of infrastructure at the system level, integrating energy efficiency, carbon intensity, water consumption, renewable energy share, heat recovery, embodied carbon, and utilization, thereby overcoming the limitations of metrics defined at the level of individual facilities. The *CHS* confederal score breaks down the comparative assessment into three distinct components—relating, respectively, to computational capacity, system properties, and suitability for artificial intelligence workloads—because reducing it to a single index would obscure the very trade-offs that the assessment aims to highlight. These are complemented by the federation gain,

$$G_F = \mathrm{CHS}(\mathcal{F}) / \sum_i \mathrm{CHS}(n_i),$$

which measures how much cooperation adds beyond the simple sum of its parts and whose value, as this volume will show, is not always greater than one.

I.5 Structure and Organization of the Book: The Eighteen Parts and Their Argumentative Logic

The work is divided into eighteen Parts and sixty-five chapters, followed by the Conclusions and a set of appendices containing the formal models, the regulatory framework, and the territorial atlases. The sequence of the Parts is not thematic but argumentative: each builds upon the preceding ones and sets the stage for the subsequent ones, so that reading the book sequentially reconstructs the development of the thesis in the order in which it was elaborated.

Part One establishes the problem. It demonstrates why the growth trajectory of computational demand encounters material and institutional constraints that scale concentration is unable to absorb; it traces the history of the oscillation between centralization and distribution to position the confederation as the fourth stage of evolution; and it distinguishes the notion of physical scale from that of logical scale, ultimately formulating the definition of a confederal data center.

Part Two constructs the theory. Here, the confederation is derived from the framework of cooperative digital infrastructures; its structure is formalized; its architectural stack is extended to the geographic scale; its laws of scale are studied; and its two key metrics—the coupling index and effectively cooperative computation—are defined. This is the most densely structured part from a formal standpoint, and it is the one to which all subsequent parts refer.

The third and fourth parts apply the theory. The third Part analytically describes the four classes of nodes—ranging from a proximity node of about one hundred kilowatts to a one-gigawatt campus—and compares their combinations at equal installed power. The fourth

Part constructs a national reference scenario, quantifies its capacity, and systematically compares it with the monolithic alternative, showing under which conditions each configuration prevails.

Parts 5 through 9 address the physical dimensions of the architecture. Computational geography transforms site selection into the simultaneous selection of a network of sites, using the mathematical framework of localization. The federated energy infrastructure treats the confederation as a constituent of the electric system rather than merely a user of it. Thermal federation addresses cooling as a problem whose nature changes with scale and heat recovery as a territorial service. The network fabric constructs the latency and bandwidth balances that make cooperation possible. Finally, confederated computing and storage define the level that makes the confederation usable as a single logical machine, subject to the constraints of data locality and sovereignty.

Parts 10 through 14 address the system's operation. Distributed artificial intelligence maps families of techniques to node classes and identifies thresholds beyond which a load cannot be distributed. The confederation's operating system describes orchestration, decision-making under uncertainty, and digital twins. Resilience addresses cooperative redundancy, failure domains, and native continuity. Security addresses zero-trust, distributed consensus resilient to quantum threats, and structurally private storage. Systemic sustainability establishes a system metric, life-cycle assessment, and water constraints.

The final four parts conclude the discussion. The economics of the confederation compares capital and operating costs across classes and introduces the notion of federation economies alongside those of scale and scope. The part dedicated to governance, territory, and society analyzes ownership structures, the role of the territorial digital prosumer, and the implications for industrial policy. The European section extends the model to a confederation of national confederations and quantifies the resulting gains. The final section tests the thesis through a systematic comparison of canonical configurations, the construction of a dominance map, and the formulation of the optimal scale theorem along with its conditions of validity.

The volume concludes with appendices, which serve a substantive rather than decorative purpose. They contain the mathematical models in full form, the mathematics of reliability, coupling models, energy models, thermal models, network models, and cost models, the European and national regulatory frameworks, the simulation methodology, reference architectures ranging from the order of hundreds of kilowatts to gigawatts, and the Italian and European territorial atlases. Finally, a , structural convention runs through all the chapters: each ends with a section that applies its topic to a systematic comparison between large-scale, regional, territorial, and local nodes and discusses under what conditions the confederation adds value relative to each of them. This recurring structure is intentional, as it allows the thesis to be tested across every dimension of the problem rather than only in the concluding section.