GREEN COMPUTING

Dr. U. Hemamalini

Dr. S. Arunarani

Dr. A. Bharathi

Dr.N.Shyamala devi

Imaginex Inks Publication

71A 71B First Street, RKV Avenue,
Old Pallavaram, Chennai 600117, India.
Phone: +919962991087
e-mail: info@imaginexinkspublication.com
https://www.imaginexinkspublication.com/

GREEN COMPUTING

Authored by

Dr. U. Hemamalini. Dr. S. Arunarani, Dr. A. Bharathi and Dr.N.Shyamala devi

14th May, 2025

Price: Rs. 450/-

ISBN: 978-93-47966-00-2

Published by and copies can be had from:

Imaginex Inks Publication

71A 71B First Street, RKV Avenue,

Old Pallavaram, Chennai 600117, India.

Phone:9750663871, 9962991057

e-mail: info@imaginexinkspublication.com

https://www.imaginexinkspublication.com/

Preface

The digital infrastructure of the twenty-first century has an environmental cost that often remains invisible: rising electricity use, growing electronic waste, and increasing carbon emissions. With the rapid expansion of artificial intelligence, cloud computing, 5G networks, and large-scale data centres, the need for sustainable digital practices has become urgent.

This book introduces the principles and practices of green computing with special attention to India, where digital infrastructure is expanding rapidly. It explains how engineering choices, software design, hardware efficiency, data centre management, and policy decisions can reduce the environmental impact of computing systems.

Organised into four parts, the book covers the foundations of computation, sustainable hardware and data centres, green software and artificial intelligence, and measurement frameworks for net-zero computing. It is intended for students, researchers, engineers, and policymakers seeking a clear understanding of sustainable computing in a changing technological world.

Acknowledgments

We express our sincere gratitude to **Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai**, for providing the academic environment, encouragement, and support for the preparation of this book.

We are deeply thankful to the management, authorities, and academic leadership of VISTAS for their continuous motivation and guidance. We also extend our heartfelt thanks to our colleagues, faculty members, research scholars, and students for their valuable suggestions, cooperation, and encouragement throughout the development of this work.

We are grateful to our families and well-wishers for their constant support and inspiration. Finally, we thank the publisher for their assistance and cooperation in bringing out this book successfully.

TABLE OF CONTENTS

PART III · SOFTWARE, AI, AND CLOUD

GREEN COMPUTING

Dr. U. Hemamalini

Dr. S. Arunarani

Dr. A. Bharathi

Dr.N.Shyamala devi

Imaginex Inks Publication

71A 71B First Street, RKV Avenue,
Old Pallavaram, Chennai 600117, India.
Phone: +919962991087
e-mail: info@imaginexinkspublication.com
https://www.imaginexinkspublication.com/

GREEN COMPUTING

Authored by

Dr. U. Hemamalini. Dr. S. Arunarani, Dr. A. Bharathi and Dr.N.Shyamala devi

14th May, 2025

Price: Rs. 450/-

ISBN: 978-93-47966-00-2

Published by and copies can be had from:

Imaginex Inks Publication

71A 71B First Street, RKV Avenue,

Old Pallavaram, Chennai 600117, India.

Phone:9750663871, 9962991057

e-mail: info@imaginexinkspublication.com

https://www.imaginexinkspublication.com/

Preface

The digital infrastructure of the twenty-first century has an environmental cost that often remains invisible: rising electricity use, growing electronic waste, and increasing carbon emissions. With the rapid expansion of artificial intelligence, cloud computing, 5G networks, and large-scale data centres, the need for sustainable digital practices has become urgent.

This book introduces the principles and practices of green computing with special attention to India, where digital infrastructure is expanding rapidly. It explains how engineering choices, software design, hardware efficiency, data centre management, and policy decisions can reduce the environmental impact of computing systems.

Organised into four parts, the book covers the foundations of computation, sustainable hardware and data centres, green software and artificial intelligence, and measurement frameworks for net-zero computing. It is intended for students, researchers, engineers, and policymakers seeking a clear understanding of sustainable computing in a changing technological world.

Acknowledgments

We express our sincere gratitude to **Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai**, for providing the academic environment, encouragement, and support for the preparation of this book.

We are deeply thankful to the management, authorities, and academic leadership of VISTAS for their continuous motivation and guidance. We also extend our heartfelt thanks to our colleagues, faculty members, research scholars, and students for their valuable suggestions, cooperation, and encouragement throughout the development of this work.

We are grateful to our families and well-wishers for their constant support and inspiration. Finally, we thank the publisher for their assistance and cooperation in bringing out this book successfully.

TABLE OF CONTENTS

The Digital Planet’s Hidden Weight

Why Computing Must Go Green

Chapter 1

The Digital Planet's Hidden Weight

Let the reader consider the following scenario, which is by no means a hypothetical one. In Navi Mumbai, within the Yotta D1 facility — one of the largest data centres on the Indian subcontinent — row upon row of server racks hum continuously, twenty-four hours a day, three hundred and sixty-five days a year. Outside, the summer temperature climbs above 38°C. Inside the facility, sophisticated cooling infrastructure fights a quiet, relentless battle against the heat that the servers themselves produce. The electricity consumed by this single campus in a single year is comparable to the annual energy requirement of a moderately sized Indian city. And this is but one facility, in one country, on a planet that now hosts tens of thousands of such installations. The question that the present book places before the reader is not merely a technical one, though it has technical dimensions of the deepest kind. It is, at its root, a question of responsibility: what obligations do engineers, computer scientists, policy makers, and citizens bear toward the planet that sustains the digital infrastructure they have built? The answer, it will be argued in the chapters that follow, demands both scientific rigour and moral seriousness in equal measure.

1.1 Setting the Context: The Energy Appetite of Modern Computing

It is pertinent to begin the present discussion by establishing, with reasonable precision, the scale of the problem under consideration. The information and communications technology (ICT) sector has,

over the past three decades, grown from a relatively modest consumer of global electricity into one of the most significant industrial energy users on the planet. To appreciate this transformation fully, it is necessary to consider not merely the raw consumption figures — significant as they are — but also the trajectory along which those figures are moving, and the forces that are driving that trajectory.

As per the estimates published by the International Energy Agency in its 2024 review, data centres globally consumed between 240 and 340 terawatt-hours (TWh) of electricity in the year 2022. To place this figure in a context more immediately familiar to the Indian reader: this is roughly equivalent to three times the total electricity generated by all of India's hydropower plants in a comparable period. Data transmission networks — the cables, routers, and switching infrastructure that carry digital information across the globe — add a further 200 to 270 TWh annually. When one accounts for end-user devices, the manufacturing of hardware, and the energy embedded in the built infrastructure of the digital economy, the total ICT energy footprint rises to somewhere in the range of 1,500 to 2,500 TWh per year, depending on the methodological choices made in the accounting exercise.

The Indian context deserves particular and careful attention in this regard, as it illustrates both the promise and the peril of digital growth with unusual clarity. India's data centre capacity has grown at a compound annual rate exceeding 20 per cent in recent years, driven by the rapid digitalisation of the economy, the explosive adoption of smartphones, the growth of cloud services, and government initiatives such as Digital India. The Bureau of Energy Efficiency (BEE), operating under the Ministry of Power, has estimated that data centres in India collectively consumed

approximately 6.5 to 7 billion units (kilowatt-hours) of electricity in 2022, a figure that is projected to grow substantially as hyperscale facilities come online in Mumbai, Chennai, Hyderabad, and Pune. At the same time, India has committed under its Nationally Determined Contributions (NDCs) to reduce the emissions intensity of its GDP by 45 per cent from 2005 levels by 2030. The tension between these two realities — rapidly growing digital energy demand and ambitious decarbonisation commitments — is not merely a policy problem. It is an engineering challenge of the first order, and it is precisely this challenge that the discipline of green computing is called upon to address.

It would be a mistake, however, to regard this as a problem unique to developing economies. The developed world faces an equally significant, if differently shaped, version of the same challenge. The United States alone hosts approximately 33 per cent of the world's data centre capacity. The hyperscale facilities operated by the major cloud providers — Amazon Web Services, Microsoft Azure, Google Cloud — each consume electricity in quantities that rival the consumption of small nations. The emergence of large-scale artificial intelligence training workloads has added a new and particularly energy-intensive dimension to this picture: training a single large language model of the scale of GPT-3 was estimated to have generated approximately 552 metric tonnes of carbon dioxide equivalent — a figure roughly comparable to the lifetime carbon footprint of five average American automobiles.

The present chapter does not seek to overwhelm the reader with statistics, of which there is no shortage, but rather to equip the reader with a conceptual framework within which those statistics can be interpreted critically and usefully. To that end, the sections

that follow examine the principal dimensions of computing's environmental footprint in turn: energy consumption and its carbon implications, the rebound effect that undermines naive efficiency optimism, electronic waste and the materials dimension of the problem, supply chain emissions, and the less-discussed but increasingly critical question of water consumption.

Table 1.1 Global ICT Energy Consumption — Sectoral Breakdown (2022)

ICT Sector	Estimated Consumption (TWh)	Share of ICT Total	% of Global Electricity	Trend (2020–2025)
Data centres (hyperscale + co-location + enterprise)	240–340	~18%	0.9–1.3%	Accelerating — AI workloads primary driver
Data transmission networks (core + metro + access)	200–270	~14%	0.8–1.0%	Moderate — efficiency partially offsetting traffic growth
End-user devices (smartphones, laptops, televisions)	600–700	~38%	2.3–2.7%	Slow — device efficiency improving with each generation
Customer premises network equipment (routers, modems)	150–200	~11%	0.6–0.8%	Slow growth; modest efficiency improvement
Hardware manufacturing (embodied/embedded energy)	300–400	~19%	1.2–1.5%	Growing — shorter refresh cycles, AI chip demand

Total ICT (Scope 1 + 2, conservative estimate)	~1,500	100%	~5.8%	—
Total ICT (inclusive of full Scope 3 supply chain)	~2,100–2,500	—	~8–10%	—

Sources: International Energy Agency (2024); Masanet et al. (2020); Freitag et al. (2021); Malmodin and Lundén (2018); Bureau of Energy Efficiency, India (2022). The wide ranges in certain cells reflect genuine methodological disagreement among researchers rather than imprecision in measurement. Scope 3 figures include upstream supply chain activities and downstream end-of-life processing.

1.2 The Rebound Effect: A Cautionary Lesson from Economic History

Among the most important — and most frequently overlooked — concepts in the economics of energy efficiency is the rebound effect, also known, in its most extreme form, as the Jevons paradox. It is a concept that every student of green computing must understand thoroughly, because it explains a pattern that has recurred with remarkable consistency in the history of computing, and because its implications for policy and engineering strategy are profound.

The original observation was made not in the context of computing at all, but in the context of coal and steam power, by the British economist William Stanley Jevons in his 1865 work

The original observation was made not in the context of computing at all, but in the context of coal and steam power, by the British economist William Stanley Jevons in his 1865 work *The Coal Question*. Jevons noted, with some care for the evidence, that improvements in the efficiency of James Watt's steam engine — which had substantially reduced the amount of coal required to produce a given quantity of mechanical work — had not resulted

in a reduction in Britain's total coal consumption. On the contrary, coal consumption had increased dramatically, because the lower cost of steam power had made it economically viable to use that power in an expanding range of applications, thereby generating additional demand that more than offset the savings from improved efficiency. This is the essential logic of the rebound effect: efficiency improvements lower the effective cost of a service, which stimulates demand for that service, which may partially or fully offset the energy savings that the efficiency improvement was intended to achieve.

In the domain of computing, the rebound effect is not a theoretical speculation but an empirically documented historical reality. It is worth dwelling on this point, because there is a persistent and understandable tendency among engineers to assume that making something more efficient must, of necessity, reduce its total resource consumption. The history of computing contradicts this assumption at every turn. The energy efficiency of computer processors — measured in useful operations performed per joule of energy consumed — has improved by a factor of approximately one thousand over the past thirty years. Global electricity consumption by computing over the same period has grown substantially. The efficiency improvement did not reduce consumption; it enabled new applications, expanded user populations, made previously uneconomical services commercially viable, and generated demand that absorbed and exceeded the efficiency gains.

To make this concrete with an Indian example: the mobile data revolution in India, catalysed by the launch of Jio's 4G network in September 2016 and the associated dramatic reduction in the cost of mobile data, resulted in an explosion of data consumption that no energy efficiency improvement in network

infrastructure could conceivably offset in absolute terms. Monthly mobile data traffic in India, which stood at approximately 200 petabytes in 2016, had grown to over 13,000 petabytes by 2022 — a sixty-five-fold increase in six years. The energy efficiency of the network improved substantially over this period; the total energy consumed by India's mobile network grew nonetheless, because the scale of demand growth was many times larger than the scale of efficiency improvement.

The lesson that one must draw from this is not that energy efficiency is unimportant or counterproductive — it remains essential, and its absence would make the situation considerably worse. Rather, the lesson is that energy efficiency is a necessary but not sufficient condition for absolute decarbonisation. Structural measures — carbon pricing mechanisms, renewable energy obligations, product longevity requirements, carbon-aware workload scheduling, and demand-side management — are required in conjunction with efficiency improvements if the goal is an absolute reduction in emissions, rather than merely a relative reduction in emissions per unit of computational output. The distinction between absolute and relative decarbonisation is one that the present book will return to repeatedly, as it is a distinction that is systematically blurred in a great deal of corporate sustainability communication.

1.3 Electronic Waste: The Downstream Burden

The environmental conversation about computing is, for the most part, a conversation about energy. This is understandable, given that energy is the most readily measurable and most directly regulated dimension of environmental impact. It is, however, an incomplete conversation, because it leaves largely unexamined the materials dimension of the problem — the physical stuff of which

computing devices are made, the environmental cost of extracting and processing that stuff, and the consequences of its disposal at the end of the device's useful life.

Electronic waste — commonly referred to as e-waste — is the fastest-growing solid waste stream in the world. The Global E-waste Monitor 2024 reported that 62 million metric tonnes of e-waste were generated globally in 2022, representing an increase of approximately 82 per cent from the 34 million tonnes recorded in 2010. Of this total, only 22.3 per cent was collected and processed through formal, documented recycling channels. The remainder — representing approximately 48 million tonnes of material — was discarded through informal channels, landfilled, or incinerated, often in ways that release hazardous materials into the environment and expose workers and local communities to serious health risks.

India's position in this picture is significant and somewhat paradoxical. On the one hand, India is among the largest generators of e-waste in the world, ranked fifth globally in absolute terms with approximately 1.6 million metric tonnes generated in 2019, a figure that has grown with the expansion of the consumer electronics market. On the other hand, India has one of the most extensive — if largely informal — e-waste recycling ecosystems in the world, concentrated in urban centres such as Delhi, Mumbai, Bengaluru, and Chennai, where livelihoods have been built around the recovery of valuable materials from discarded electronics. The challenge for Indian policy and engineering in this context is to formalise and upgrade this informal sector, capturing its remarkable resourcefulness while eliminating the health and environmental harms that informal processing causes.

It is pertinent to note, in this context, that the environmental cost of hardware is not confined to its end-of-life phase. The manufacturing of a modern microprocessor is an extraordinarily

material-intensive process. A semiconductor fabrication facility — such as the TSMC plants in Taiwan that manufacture the chips used in virtually every major computing device in the world — consumes enormous quantities of ultrapure water, specialty chemicals, and energy. The embodied energy and carbon in a smartphone — the energy and emissions incurred in mining raw materials, refining them, fabricating components, and assembling the final device — accounts for approximately 70 to 80 per cent of the device's total lifetime carbon footprint. The remaining 20 to 30 per cent is generated during the years of use. A green computing strategy that focuses exclusively on operational efficiency while ignoring device longevity, repairability, and end-of-life processing is, in effect, addressing the smaller part of the problem while appearing to address the whole of it.

Table 1.2 Global E-Waste and Data Centre Metrics — Selected Indicators (2015–2025)

Indicator	2015	2018	2020	2022	2025 (projected)
Global e-waste generated (million tonnes)	44.4	50.0	53.6	62.0	~74
E-waste formally recycled (% of generated)	20%	20%	17.4%	22.3%	~25% (target)
India e-waste generated (million tonnes)	~1.0	~1.4	~1.6	~2.0	~3.0 (est.)
Global data centre electricity (TWh)	~200	~205	~230	~240–340	~400–500
Average hyperscale PUE (global)	1.65	1.58	1.55	1.48	~1.40

Renewable energy share — major cloud providers (%)	~20%	~35%	~50%	~65%	~80%
Global data centre direct water consumption (bn m³)	~1.7	~2.0	~2.2	~2.5	~3.0

Sources: Global E-waste Monitor (2024); IEA Data Centres Report (2024); Uptime Institute Annual Report (2024); MeitY, Government of India (2022); Masanet et al. (2020). Figures for India derived from Central Pollution Control Board (CPCB) estimates. PUE figures represent hyperscale leader averages; global fleet average PUE remains approximately 1.55–1.60, reflecting the large installed base of older, less efficient facilities.

1.4 The Supply Chain: Understanding Scope 3 Emissions

The Greenhouse Gas Protocol, which has emerged as the internationally accepted standard for corporate emissions accounting, organises emissions into three categories, or 'scopes.' Scope 1 covers direct emissions from sources owned or controlled by the reporting organisation — for example, the combustion of diesel in on-site generators or the leakage of refrigerants from cooling systems. Scope 2 covers indirect emissions from the purchase of electricity, heat, or steam. Scope 3 covers all other indirect emissions that occur in the value chain of the reporting organisation, both upstream — in the production of goods and services that the organisation purchases — and downstream — in the use and disposal of the products that the organisation sells.

It is a matter of considerable importance — and, it must be said, considerable corporate inconvenience — that Scope 3 emissions typically account for 70 to 90 per cent of the total lifecycle carbon footprint of a technology company. When Apple Inc. disclosed in its 2024 Environmental Progress Report that 70 per cent of its total carbon footprint originated in Scope 3 activities, primarily hardware manufacturing, it was making public what

analysts had long understood: the operational electricity consumption of Apple's own facilities is a small fraction of the total carbon cost of putting an Apple product into a consumer's hands. For cloud providers, the arithmetic is somewhat different — the manufacturing of servers, networking equipment, and cooling infrastructure represents a significant but variable fraction of total lifecycle footprint — but the principle remains the same: focusing on operational efficiency alone systematically underestimates the problem.

For the Indian technology industry, the Scope 3 challenge has distinctive features. India's manufacturing sector, which is growing rapidly in the electronics domain as companies seek to diversify supply chains away from China, operates with a carbon intensity that reflects India's current electricity generation mix — still substantially coal-dependent, though the share of renewables is growing encouragingly. As Indian firms take on greater roles in the global electronics supply chain, their Scope 3 emissions — which become the Scope 1 and 2 emissions of their customers — will come under increasing scrutiny from international buyers subject to mandatory supply chain disclosure requirements, including the European Union's Corporate Sustainability Reporting Directive (CSRD), which came into force in 2024.

It is pertinent to note, in this context, that the Government of India's Production Linked Incentive (PLI) scheme for electronics manufacturing represents an opportunity as well as a responsibility. The investment it attracts could be directed toward manufacturing facilities powered by renewable energy, designed for resource efficiency, and equipped with take-back systems for end-of-life products — or it could lock in another generation of carbon-intensive manufacturing infrastructure. The choices made

in the design of these facilities over the next five years will have consequences that endure for two to three decades.

1.5 Water Consumption: The Cooling Burden That Numbers Rarely Capture

Among the environmental impacts of digital infrastructure, water consumption is perhaps the least discussed and, in the context of a country like India — where freshwater stress affects hundreds of millions of people — among the most consequential. Data centres consume water through two principal pathways. The first is direct: evaporative cooling systems, which remove heat from server halls by evaporating water into the atmosphere, consume freshwater continuously and in large quantities. The second is indirect: thermoelectric power plants, which supply the majority of the world's grid electricity, consume water for steam generation and cooling, and every unit of electricity consumed by a data centre carries an embedded water cost determined by the generation technology used to produce it.

The Water Usage Effectiveness (WUE) metric — defined as the volume of water consumed, in litres, per kilowatt-hour of IT equipment energy — provides a facility-level measure of direct water consumption. A well-managed modern facility might achieve a WUE of 0.5 to 1.0 litres per kilowatt-hour; an older or less carefully managed facility might exceed 3.0 litres per kilowatt-hour. At the scale of a large data centre consuming, say, 100 megawatts of IT load, even a WUE of 1.0 represents a daily water consumption of approximately 2.4 million litres — water that is evaporated and cannot be recovered.

In the Indian context, this is a matter that demands explicit attention. Several of the cities that have emerged as preferred locations for data centre development — Navi Mumbai,

Hyderabad, Chennai, and Pune — are located in regions where freshwater availability is seasonally stressed, groundwater tables are declining, and municipal water supply is already under pressure from residential and industrial demand. The siting of large data centres in water-stressed areas, without careful assessment of cumulative water impact, represents a form of environmental externality that the industry has been slow to address. Microsoft's disclosure in 2023 that its global water consumption increased by 34 per cent year-on-year — attributable in significant part to the cooling demands of artificial intelligence workloads — brought unusual public attention to this issue, and it is one to which Indian policymakers and data centre developers would do well to attend proactively rather than reactively.

1.6 Towards a Systemic Understanding: What This Book Seeks to Accomplish

Having surveyed the principal dimensions of computing's environmental footprint — energy consumption, the rebound effect, electronic waste and material impacts, supply chain emissions, and water consumption — the reader will appreciate that the challenge of green computing is not reducible to any single technical intervention, however sophisticated. It is, rather, a systemic challenge that requires a systemic response: one that addresses hardware design, data centre engineering, network architecture, software development practices, cloud resource management, measurement and accountability frameworks, and long-term technology strategy simultaneously and in an integrated manner.

The present book is organised to address these dimensions in sequence while maintaining, throughout, the systemic perspective that their interconnection demands. Part I, of which the present

chapter forms the opening statement, establishes the foundational understanding — the scale of the problem and the physics that governs its solutions. Part II examines the major physical infrastructure components: hardware, data centres, and networks. Part III turns to the software, artificial intelligence, and cloud computing dimensions, which represent the fastest-growing sources of demand and, simultaneously, some of the most powerful levers for efficiency improvement. Part IV addresses the measurement, policy, and emerging technology dimensions that will determine whether the next decade's digital growth can be reconciled with a credible path to net-zero computing.

It is the sincere aspiration of the author that the reader — whether an undergraduate encountering these ideas for the first time, a postgraduate scholar developing a research agenda, a practising engineer seeking to make more responsible design choices, or a policy maker seeking to understand what the technology community is and is not capable of delivering — will find in the chapters that follow not only useful information but a framework for thinking that outlasts the specific numbers it contains. The numbers, as has already been noted, are moving. The framework, it is hoped, will prove more durable.

Thermodynamics, Computation, and the Physics of Energy Use

Chapter 2

Thermodynamics, Computation, and the Physics of Energy Use

In the year 1961, Rolf Landauer — a physicist working at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York — published a paper of remarkable prescience in the IBM Journal of Research and Development. The question he addressed was, on the surface, an abstract one: does the logical structure of a computation impose any fundamental limit on the energy it must consume? His answer, arrived at through careful thermodynamic reasoning, was both precise and, to many of his contemporaries, surprising. The erasure of a single bit of information — the resetting of a computational memory cell from an unknown state to a known one — must, of physical necessity, dissipate a minimum quantity of energy as heat. Not because of poor engineering. Not because of inadequate materials. Because the second law of thermodynamics demands it.

That minimum quantity, at the temperatures at which ordinary computers operate, is approximately 2.87×10^{-21} joules per bit erased. The processors manufactured today at the Indian Space Research Organisation's Semiconductor Laboratory in Chandigarh, or at the cutting-edge foundries of TSMC and Samsung that supply them, consume energy per operation that is some ten million times larger than this fundamental limit. The gap between where the present state of the art stands and where the laws of physics permit it to go is the single most important fact in the long-term engineering of green computing. It tells us, without ambiguity, that there is vastly more room for improvement than

has yet been realised, and that the work of realising it is among the most consequential engineering challenges of our time.

2.1 Landauer's Principle and the Thermodynamic Cost of Forgetting

In order to appreciate the significance of Landauer's result, it is necessary to understand the conceptual connection it establishes between two disciplines that had, until his work, developed largely independently of one another: information theory and thermodynamics. Information theory, as developed by Claude Shannon in his foundational 1948 paper, concerned itself with the mathematical structure of information — how it could be quantified, transmitted, compressed, and protected against noise — without reference to the physical medium in which it was encoded. Thermodynamics, the study of heat, work, and entropy, concerned itself with the physical behaviour of macroscopic systems without, in its classical formulation, any reference to information as such.

Landauer's insight was that these two bodies of theory were, at a fundamental level, describing the same thing from different vantage points. The information content of a physical system — the number of binary choices that would be required to fully specify its state — is directly related to its thermodynamic entropy. When a bit of information is erased — when a physical system representing either a 0 or a 1 is reset to a known, definite state — the information about what state it was in is irreversibly destroyed. This destruction of information corresponds, in thermodynamic terms, to an increase in the entropy of the system's environment. The second law of thermodynamics requires that this entropy increase be accompanied by the dissipation of heat equal to at least $kT \ln 2$, where k is Boltzmann's constant (approximately $1.38 \times$

10^{-23} joules per kelvin) and T is the absolute temperature of the environment.

At room temperature — approximately 293 kelvin — this evaluates to about 2.87×10^{-21} joules, or 2.87 zeptojoules. This is an extraordinarily small quantity of energy; for comparison, a single photon of visible light carries approximately one hundred times more energy. The Landauer limit is so far below the energy consumption of current computing hardware that it might seem irrelevant to any practical engineering discussion. It is not. It matters precisely because it establishes that the energy cost of computation is not a fixed property of the computational task being performed, but a consequence of the physical implementation of that computation — and that current implementations are vastly less efficient than the laws of physics require. The gap between present and ideal is not a small margin to be optimised away at the edges; it is a factor of ten million, and it represents the entire horizon of future improvement.

It should be noted, for completeness, that Landauer's principle applies specifically to the erasure of information — to irreversible logical operations that destroy information about prior states. Logical operations that are reversible — that is, operations from which the input can be uniquely recovered given the output — do not, in principle, incur the Landauer energy cost. This observation, elaborated by Charles Bennett in his foundational 1982 paper, forms the theoretical basis for the concept of reversible computing, which is examined in the section that follows.

The experimental verification of Landauer's principle, which had to await the development of sufficiently precise experimental apparatus, was accomplished by Bérut and colleagues in 2012. Using an optical tweezer system to manipulate a single colloidal particle in a double-well potential — a physical realisation of a

single-bit memory — they demonstrated that the erasure of one bit of information dissipated an amount of heat that matched the theoretical prediction to within experimental uncertainty. This elegant experiment put to rest any remaining doubts about the physical reality of the principle and established it firmly as a foundational result in the science of computation.

2.2 Reversible and Adiabatic Computing: Approaching the Ideal

The theoretical possibility of computation without Landauer energy cost — computation that operates through reversible logical steps and thus, in principle, need not dissipate heat at all — has attracted sustained research interest since Bennett's 1973 demonstration that any Turing-computable function can be computed reversibly. A logical operation is reversible if and only if it is bijective: that is, if a unique output corresponds to every possible input, and conversely, if the input can be uniquely recovered from the output. The standard logical gates of classical computing — AND, OR, NOT — are not all reversible in this sense: an AND gate whose output is 0 may have received any of three distinct input combinations (00, 01, or 10), and the input cannot be recovered from the output alone.

Reversible logic gates have been designed and studied extensively. The Toffoli gate (also known as the controlled-controlled-NOT gate) and the Fredkin gate (the controlled-SWAP gate) are both reversible and, taken together, are universal for reversible computation — meaning that any Boolean function can be computed using combinations of these gates alone. Reversible computation necessarily carries additional information through each step of the calculation, preserving the 'garbage' bits that

encode the history of the computation, but it avoids the Landauer cost by avoiding information erasure.

Adiabatic computing represents a physically realisable approach to approximating reversible computing in actual electronic circuits. In standard CMOS circuit operation, the switching of a logic gate involves the rapid charging or discharging of a capacitance, a process that dissipates energy at a rate proportional to the square of the supply voltage regardless of how much of that energy is actually used for the computation. In adiabatic circuit design, supply voltages are changed slowly and gradually — in a ramp or sinusoidal fashion — allowing charge to be recovered and recycled rather than dissipated as heat. The energy dissipation of an adiabatic circuit scales with the speed of the voltage ramp: the slower the ramp, the less energy is dissipated, approaching zero in the limit of infinitely slow operation. Research circuits employing adiabatic techniques have demonstrated energy dissipations several orders of magnitude below those of equivalent conventional CMOS circuits, at the cost of reduced operating speed.

It must be acknowledged, however, that reversible and adiabatic computing remain, as of the date of writing, primarily research topics rather than commercially deployed technologies, with the exception of certain specialised applications in low-power IoT devices and cryptographic hardware. The practical challenges of manufacturing adiabatic circuits at scale, designing programming models that map efficiently to reversible logic, and managing the additional complexity of garbage bit handling have thus far limited their adoption. Quantum computing, whose logical operations are inherently unitary — and therefore reversible — offers another avenue toward low-dissipation computation, but the substantial cooling requirements of current quantum hardware

place its net energy costs far above those of classical processors for most practical workloads. Both directions are examined further in Chapter 10 of the present volume.

2.3 Dynamic Voltage and Frequency Scaling: The Engineer's Primary Lever

Having established the theoretical bounds on computation's energy cost, the present section turns to the most practically significant mechanism through which the energy consumption of real processors is managed: Dynamic Voltage and Frequency Scaling, universally abbreviated as DVFS. This technique is the primary means by which modern processors adapt their power consumption to the computational demand placed upon them, and an understanding of its physical basis is essential for any serious practitioner of green computing.

The dynamic power consumed by a CMOS integrated circuit — the power consumed by transistors actively switching from one logical state to another — is governed by the relation $P = \alpha C f V^2$, in which α represents the activity factor (the proportion of transistors switching on each clock cycle), C represents the total switching capacitance of the circuit, f represents the operating frequency, and V represents the supply voltage. Of the quantities in this expression, the supply voltage is by far the most influential: power scales with the square of voltage, meaning that a reduction in supply voltage by a factor of two reduces dynamic power by a factor of four. Since, in a well-designed CMOS circuit, the maximum operating frequency scales approximately linearly with supply voltage (above the transistor threshold), reducing voltage simultaneously reduces frequency — and the net result is that energy per operation, which is proportional to P/f, scales as V^2, improving by a factor of four for a halving of voltage.

In practical terms, DVFS enables a processor to operate at a spectrum of voltage-frequency combinations — known as operating points or P-states — each offering a different trade-off between computational throughput and energy consumption. When the workload demand is high — during a complex computation, a video encoding task, or a demanding AI inference request — the processor operates at its highest P-state, maximising throughput at the cost of higher power. When demand is light — during periods of inactivity, background housekeeping tasks, or low-priority processing — the processor transitions to a lower P-state, sacrificing some throughput for substantially reduced power consumption. The operating system, in cooperation with hardware controllers, manages these transitions continuously, adapting to workload changes on timescales of milliseconds.

The reader who is familiar with the power management behaviour of modern smartphones will recognise this principle in practice. The ARM big.LITTLE architecture — which combines high-performance processor cores with smaller, more energy-efficient cores on the same chip — is, in effect, an architectural extension of the DVFS principle. Demanding tasks are routed to the high-performance cores operating at high voltage and frequency; routine tasks are handled by the efficiency cores at much lower energy cost. This architectural approach, pioneered in the mobile domain and now adopted across the computing spectrum from microcontrollers to server processors, represents one of the most significant practical advances in green processor design of the past decade.

Table 2.1 Landauer Limit vs. Actual Processor Energy per Operation (Room Temperature, 2024)

Processor / Technology Node	Energy per Operation	Multiple Above Landauer Limit	Primary Dissipation Mechanism	Principal Limitation
Landauer limit (thermodynamic minimum, 293 K)	2.87×10^{-21} J (2.87 zJ)	1× (reference)	Irreversible bit erasure — unavoidable	Fundamental thermodynamic law
Research adiabatic CMOS circuit	~1×10^{-15} J (1 fJ)	~350,000×	Residual non-adiabatic losses; charge recycling overhead	Speed severely limited; not manufacturable at scale
ARM Cortex-M0+ (IoT, 45 nm)	~10×10^{-12} J (10 pJ)	~3.5×10^{9}×	Capacitive switching at reduced voltage; leakage at low supply	Limited performance; suitable for IoT sensors only
ARM Cortex-A76 (7 nm, mobile)	~0.5×10^{-12} J (0.5 pJ)	~1.7×10^{8}×	DVFS-managed switching; near-threshold leakage growth	Thermal density limits sustained peak performance
Apple M3 (3 nm, laptop/desktop)	~1×10^{-12} J (1 pJ)	~3.5×10^{8}×	Efficient big.LITTLE architecture; NPU offloading	Advanced packaging cost; limited repairability
NVIDIA H100 GPU (4 nm, AI training)	~50×10^{-12} J (50 pJ per FLOP)	~1.7×10^{10}×	High-density parallel arithmetic; HBM access dominant	Memory bandwidth bottleneck; requires liquid cooling

IBM z16 mainframe (5 nm)	~2×10^{-12} J (2 pJ)	~7×10^{8}×	Reliability and RAS overhead; balanced I/O energy	Optimised for transaction throughput, not energy per FLOP

Energy-per-operation values represent approximate chip-level energy divided by estimated operations per second at representative workload; direct cross-architecture comparison requires caution. Landauer limit at 293 K per Landauer (1961) and Bérut et al. (2012). Sources: Horowitz (2014); Theis and Wong (2017); manufacturer technical documentation; ISRO Semiconductor Laboratory technical publications.

2.4 Static Power, Dark Silicon, and Life After Dennard Scaling

For approximately four decades following the publication of Robert Dennard's foundational 1974 paper, the semiconductor industry enjoyed a period of almost magical progress in which each new generation of transistors was simultaneously smaller, faster, and more energy-efficient than its predecessor. The physical principle underlying this progress — now universally referred to as Dennard scaling — held that as transistors shrank, their operating voltages could be reduced proportionally, their switching speed would increase, and the power density of the chip would remain approximately constant. This meant that each new process generation delivered more computational performance for the same amount of energy, and at the same power budget. Moore's Law and Dennard scaling together constituted an extraordinary compounding advantage: more transistors per chip, each one faster and more efficient.

Dennard scaling effectively ceased to operate as a reliable guide to semiconductor progress in the years around 2005 to 2007. The fundamental cause was the growing importance of leakage currents — currents that flow through transistors even when they

are nominally switched off — as transistor dimensions shrank into the sub-100-nanometre regime. Below approximately 90 nanometres, the electric field across the gate oxide of a transistor became sufficiently intense that electrons could tunnel quantum-mechanically through the oxide, creating a gate leakage current essentially independent of the transistor's switching state. This static leakage power, unlike dynamic switching power, does not decrease when the processor is running light workloads or when DVFS reduces the operating frequency; it flows continuously so long as the supply voltage is applied. As transistor densities continued to increase through subsequent process generations, the aggregate leakage power of a chip grew to the point where it now represents 20 to 40 per cent of total chip power in leading-edge designs.

The consequence of this transition is what researchers have termed the dark silicon problem. If all the transistors on a modern processor die were to be operated simultaneously at maximum voltage and frequency, the power density would generate heat far in excess of what any practical cooling system could remove — the chip would sustain permanent thermal damage within seconds. Processor designers have responded by accepting that only a fraction of the transistors on a chip can be simultaneously active at full performance, with the remainder idled, power-gated, or operated at reduced voltage. This inactive fraction — transistors present on the die but deliberately kept 'dark' to manage power — is the eponymous dark silicon.

The dark silicon constraint has profoundly reshaped processor architecture over the past fifteen years, and this reshaping has significant implications for green computing. When not all transistors can be active simultaneously, the question for the chip architect shifts from 'how many general-purpose cores can be

fitted?' to 'how should the transistor budget be allocated to maximise useful work per watt for the workloads that matter?' The answer that the industry has converged upon is heterogeneous computing: chips that contain multiple different types of processing element, each specialised for a particular class of workload, so that any given task can be assigned to the element best suited to performing it efficiently. A modern mobile SoC of the kind designed at Qualcomm for Android smartphones — or the custom SoC developed by Apple for its iPhone and Mac product lines — may contain dedicated cores for high-performance computation, efficiency cores for background tasks, a neural processing unit for machine learning inference, a digital signal processor for audio and sensor processing, a video encoder and decoder, and an image signal processor, all on a single die. Routing the right work to the right processing element is, simultaneously, a performance optimisation and an energy optimisation.

The importance of this architectural evolution for the Indian computing ecosystem deserves specific mention. The Indian government's initiative to develop domestic semiconductor design and, eventually, fabrication capabilities — exemplified by the PLI scheme for semiconductors and the establishment of the India Semiconductor Mission — will succeed or fail, in significant measure, on the basis of whether the designed chips reflect the current state of the art in heterogeneous, energy-efficient architecture. The Department of Science and Technology's funding of advanced VLSI research at institutions such as the Indian Institute of Science, IIT Bombay, IIT Madras, and IIT Delhi is building the human capital that this endeavour requires, but the integration of green computing principles into the design education and research agenda at these institutions is a priority that merits deliberate attention.

Table 2.2 DVFS Operating Points — Energy and Performance Trade-offs (Illustrative Model)

Operating Mode	Supply Voltage (relative)	Clock Frequency (relative)	Dynamic Power (relative)	Throughput (relative)	Energy per Operation (relative)
Maximum (turbo/boost)	1.10×	1.10×	~1.33× (less efficient)	1.10×	~1.21×
Nominal rated operating point	1.00×	1.00×	1.00×	1.00×	1.00×
DVFS step down (voltage −25%)	0.75×	~0.75×	~0.42×	~0.75×	~0.56× (better)
DVFS step down (voltage −50%)	0.50×	~0.50×	~0.13×	~0.50×	~0.25× (substantially better)
Near-threshold voltage operation	~0.30×	~0.10×	~0.03× (dynamic only)	~0.10×	~0.30× (leakage now dominant)
Sleep / power-gated retention state	~0.20×	0×	~0.01× (leakage floor)	0×	— (no useful work performed)

This table presents a simplified pedagogical model. In practice, the leakage current increases non-linearly at low voltages, and the energy-per-operation minimum typically occurs at an intermediate voltage where the marginal reduction in dynamic power exceeds the marginal increase in relative leakage contribution. Model adapted from Chandrakasan et al. (1992) and Hennessy and Patterson (2019).

2.5 The Memory Wall: When Data Movement Dominates Energy

A student who has studied the energy consumption of processors in isolation — focusing on processor frequency, voltage, and transistor switching — will, upon encountering real computational workloads, discover a disconcerting fact: for a large and growing class of applications, the energy consumed in moving data between memory and processor dwarfs the energy consumed in performing arithmetic operations on that data. This phenomenon — sometimes called the memory wall, a term introduced by Wulf and McKee in their influential 1995 paper — is one of the most important energy considerations in modern computing, and one that the green computing practitioner cannot afford to overlook.

The fundamental cause of the memory wall is the growing disparity between the speed of processor arithmetic and the speed of memory access. Over the past three decades, processor arithmetic throughput has increased by approximately a factor of ten thousand, while DRAM memory latency has improved by only a factor of approximately five to ten. This disparity has been partially managed through the introduction of multiple levels of cache memory — fast, on-chip SRAM that stores recently accessed data close to the processing elements — but caches are effective only when the access pattern of the computation exhibits sufficient locality. For the large, irregular access patterns characteristic of graph analytics, scientific simulation, and, critically, the attention mechanisms of large neural networks, cache miss rates are high, and the processor spends substantial fractions of its time waiting for data to arrive from DRAM.

The energy dimension of this problem is even more striking than the performance dimension. Moving a single byte of data from off-chip DRAM to a processor register consumes approximately

100 to 1,000 times the energy of performing an arithmetic operation on that byte once it has arrived. For artificial intelligence inference workloads — the class of computation that has emerged as the dominant driver of data centre energy growth in the 2020s — the transformer architecture's attention mechanism requires accessing weight matrices that may contain hundreds of billions of parameters, billions of times per inference request. The energy consumed in supplying those weights to the compute units from DRAM is the single largest component of inference energy, not the arithmetic performed by the matrix multiplication units.

This analysis has immediate practical implications for green computing strategy that are not captured by processor-centric efficiency metrics. Near-memory processing architectures — which place computing logic in close physical proximity to or inside the memory array — offer potentially transformative energy savings for memory-bound workloads by reducing the distance and therefore the energy cost of data movement. High Bandwidth Memory (HBM) technology, which stacks DRAM dies directly atop the processor package using through-silicon vias, is already deployed in leading AI accelerators — including the NVIDIA H100 and AMD MI300 — precisely because its dramatically higher bandwidth per unit of energy makes it better suited to memory-bound AI workloads than conventional DDR DRAM. Processing-in-memory (PIM) architectures, which perform certain operations directly within the memory array without transferring data to a separate processor at all, represent the logical extension of this principle and are an active area of research at institutions including Samsung Research, SK Hynix, and several IIT research groups.

The practical implication for software developers — a point that the present book will develop further in Chapter 6, on green

software engineering — is that the energy cost of an algorithm cannot be assessed from its computational complexity alone. An algorithm with higher asymptotic computational complexity but better memory access locality may consume substantially less energy than a more computationally efficient algorithm with poor locality, because the dominant cost is memory access rather than arithmetic. This is a counterintuitive result by the standards of traditional algorithm analysis, and it illustrates, more broadly, the theme with which the present chapter concludes: the physical realities of semiconductor hardware impose constraints and opportunities on software design that no amount of abstraction can fully conceal.

2.6 Implications for the Practitioner: What Physics Teaches the Engineer

The reader who has followed the argument of the present chapter to this point will, it is hoped, appreciate the significance of what has been established. Landauer's principle tells us that there is no thermodynamic obstacle to improving the energy efficiency of computation by many orders of magnitude from where it stands today. The gap between the theoretical minimum and the practical state of the art is so vast that it should be a source of genuine engineering ambition rather than complacency. The failure to close this gap more rapidly is not, in the main, a consequence of physical limits; it is a consequence of the economic and institutional structures that govern the pace and direction of semiconductor research and development.

The end of Dennard scaling tells us that the path to further improvement runs through specialisation and heterogeneity rather than brute-force scaling. General-purpose processors operating at ever-higher clock frequencies are not the answer; targeted

processing elements designed for specific workload classes, managed by intelligent scheduling systems that route each task to the most energy-efficient element capable of performing it, are the engineering direction that the next decade demands. This is already understood and actively pursued at the leading semiconductor companies; it needs to become equally well understood at the level of computer science education, where algorithmic thinking has traditionally been decoupled from hardware energy awareness.

The memory wall tells us that energy efficiency in computing is not solely, or even primarily, a hardware problem. The data structure choices made by a programmer, the access patterns generated by an algorithm, the granularity at which parallelism is exploited — all of these software-level decisions have energy consequences that are as significant as the hardware characteristics of the processor on which they run. A student of green computing who understands only the hardware dimensions of the problem is equipped to address, at most, half of it. The present chapter has laid the physical foundations; the chapters that follow will build upon them, from the architecture of the data centre through the craft of the software developer to the policy frameworks within which both must ultimately operate.

Green Hardware

Energy-Efficient Processors, Memory, and Storage

Chapter 3

Green Hardware — Energy-Efficient Processors, Memory, and Storage

Consider the following observation, which is by no means obvious until one pauses to reflect upon it carefully. When the Defence Research and Development Organisation places an order for high-performance computing hardware for its modelling and simulation programmes, or when the Indian Railways commissions servers for its reservation and operations management infrastructure, the energy consumption of those machines over their operational lifetime — typically five to seven years — will cost more, in absolute rupee terms, than the purchase price of the hardware itself. In many large institutional deployments, the total cost of ownership of computing equipment is dominated not by the capital expenditure of acquisition but by the operating expenditure of electricity. The choice of hardware, in other words, is an energy decision masquerading as a procurement decision. It is the purpose of the present chapter to equip the reader with the scientific and engineering understanding needed to make such decisions wisely.

3.1 The Hardware-Energy Relationship: Why Architecture Choices Cascade

It is a common misconception, particularly among those approaching computing from a software background, that hardware energy consumption is a fixed parameter — a given property of the devices one purchases — rather than a design variable that can be influenced through informed choices at

multiple levels of the system stack. This misconception, though understandable, is consequential. The architectural choices made by hardware designers, and the procurement and deployment choices made by the organisations that operate that hardware, have energy implications that extend over years and aggregate across thousands of devices into figures of considerable magnitude.

The present chapter examines the principal hardware subsystems — processor architectures, memory hierarchies, and storage technologies — through the lens of energy efficiency. Each subsystem contributes to the total energy consumed by a computing system, and each offers specific opportunities for reduction through design innovation and informed selection. It is pertinent to note, however, that these subsystems do not operate in isolation: the energy profile of a processor is inseparable from the memory system it relies upon, and the memory system's energy is shaped by the access patterns generated by the software running on the processor. A systemic view is therefore necessary, and the present chapter endeavours to maintain that view even while examining individual components in some depth.

The economic dimension of hardware energy efficiency has become, in the second decade of the twenty-first century, increasingly prominent in procurement decisions at Indian public sector organisations. The Central Public Works Department's revised energy efficiency guidelines for government data centres, issued in 2021, and the Bureau of Energy Efficiency's Star Rating Programme for servers and storage equipment, represent institutional acknowledgement that hardware energy efficiency is a procurement criterion of the first order — not an environmental consideration to be weighed separately from cost, but a direct contributor to the total cost of ownership over the asset's operational life.

3.2 Heterogeneous Computing: The Architecture of Specialisation

The most significant architectural trend in processor design over the past decade — and the one with the most far-reaching implications for energy efficiency — is the shift toward heterogeneous computing: the integration of multiple distinct processing elements on a single chip or package, each optimised for a different class of computation. As was discussed in Chapter 2 in the context of dark silicon, the end of Dennard scaling has made it impractical to simply add more general-purpose processor cores and expect commensurate performance gains without proportionate increases in power consumption. The architectural response has been to replace some general-purpose capacity with specialised processing elements that perform particular operations far more efficiently.

The energy efficiency advantage of a specialised processor over a general-purpose one for a matched workload is substantial. A neural processing unit (NPU) executing a matrix multiplication — the dominant arithmetic operation in neural network inference — consumes approximately 10 to 100 times less energy per operation than a general-purpose CPU performing the same computation. A hardware video decoder, integrated into a mobile SoC, can decode a 4K video stream at a fraction of the energy cost of software decoding on a CPU. A cryptographic accelerator can perform AES encryption with energy efficiency that no software implementation on a general-purpose core can approach. In each case, the efficiency advantage arises from the same source: specialised hardware can be designed to implement a specific computational pattern in the minimal physical arrangement of transistors required, without the overhead of the instruction fetch,

decode, and general register management that a programmable processor requires.

For the Indian technology ecosystem, the implications of this architectural shift are significant on several levels. The fabless chip design companies that have emerged in India — including Incore Semiconductors, Signalchip (acquired by VVDN Technologies), and the several design centres established in Bengaluru and Hyderabad by companies including Qualcomm, Intel, Texas Instruments, and Nvidia — are actively engaged in the design of specialised processing elements for communications, automotive, and AI applications. The successful design of energy-efficient specialised processors is not merely a commercial opportunity; it is a contribution to the sustainability of every system into which those processors are eventually deployed.

Needless to say, the benefits of heterogeneous computing are realised only when workloads are correctly routed to the processing element best suited to handling them. A neural network inference request that is executed on a GPU when an NPU is available, or a video stream that is decoded in software when a hardware decoder is present and idle, represents a missed efficiency opportunity. The scheduling and orchestration software that manages workload routing is therefore as important to the energy efficiency of a heterogeneous system as the hardware itself — a point that the present volume will develop in Chapter 6, in the context of green software engineering.

Table 3.1 Energy Comparison of Processor Types per Operation (Representative Values, 2024)

Processing Element	Representative Application	Energy per Operation	Efficiency vs. General CPU	Thermal Design Power	Primary Limitation
General-purpose CPU core (x86, ARM)	All workloads	~1–10 pJ	1× (reference)	5–350 W (mobile to server)	High versatility cost; no specialisation benefit
GPU compute core (CUDA, ROCm)	Parallel math, AI training	~5–50 pJ per FLOP	1–2× better for parallel workloads	50–700 W	High idle power; poor efficiency at low utilisation
Neural Processing Unit (NPU/DSP)	Neural network inference	~0.1–1 pJ per MAC	10–100× better for inference	0.5–10 W	Fixed-function; limited to specific model architectures
Field Programmable Gate Array (FPGA)	Custom compute patterns	~0.5–5 pJ per operation	5–20× better for matched workloads	5–100 W	Programming complexity; lower raw performance than ASIC
Application-Specific IC (ASIC)	Single specific workload	~0.05–0.5 pJ	100–1,000× better than CPU for target	1–30 W	Inflexible; expensive NRE; long development cycle
Microcontroller (IoT class, ARM Cortex-M)	Sensing, control, edge inference	~10–100 pJ	0.1–1× (lower clock, lower precision)	0.001–1 W	Limited performance; suitable for simple tasks only

MAC = Multiply-Accumulate operation, the fundamental operation in neural network inference. NRE = Non-Recurring Engineering cost. Values are representative orders of magnitude; actual figures vary significantly by specific device, operating

frequency, workload, and measurement methodology. Sources: Horowitz (2014); Shalf (2020); manufacturer datasheets.

3.3 Memory Technology: Balancing Capacity, Speed, and Energy

The memory hierarchy of a modern computing system — from the small, fast, on-chip SRAM caches adjacent to the processor, through the larger but slower DRAM main memory, to the still larger and non-volatile storage of solid-state drives and hard disk drives — represents a set of engineering trade-offs between capacity, access speed, and energy consumption that no single technology can optimally satisfy for all workloads.

As was established in Chapter 2, the energy cost of accessing DRAM is approximately 100 to 1,000 times the energy cost of performing an arithmetic operation on data once it has arrived at the processor. This asymmetry makes the depth of the memory hierarchy — specifically, how frequently data must be fetched from DRAM rather than served from cache — one of the most important determinants of total system energy for memory-intensive workloads. Cache-friendly algorithms and data structures, which organise computation to maximise data reuse within the cache before it is evicted, are therefore a form of green computing practice, and their importance is proportional to the memory-intensity of the workload.

High Bandwidth Memory (HBM) represents the most significant recent advance in energy-efficient main memory technology for high-performance computing applications. In conventional DDR DRAM configurations, the memory chips are mounted on a printed circuit board alongside the processor, connected by a bus with limited bandwidth and relatively high energy cost per bit transferred. HBM eliminates the long

interconnect by stacking DRAM dies directly on top of — or alongside, in a 2.5D configuration — the processor die, using microscopic copper pillars called through-silicon vias (TSVs) to connect the layers. The result is a memory interface that is simultaneously wider in bandwidth and lower in energy per bit than conventional DRAM by a significant margin. The NVIDIA H100 GPU, which uses HBM3 memory, achieves memory bandwidth of approximately 3.35 terabytes per second — roughly three times the bandwidth of the best conventional GDDR memory, at lower energy per bit. For AI training and inference workloads that are bottlenecked by memory bandwidth, this improvement translates directly into energy savings per unit of useful computation performed.

It is pertinent to note, in the context of the present discussion, that 3D-stacked memory technologies carry their own thermal challenges. Stacking memory dies atop a processor die concentrates heat in a small volume, creating thermal management problems that must be addressed through careful packaging design and, in some cases, direct liquid cooling of the package. The interaction between memory architecture choices and data centre cooling requirements is therefore a system-level concern that spans the hardware and facility dimensions simultaneously — an illustration of the systemic interdependencies that the present volume aims to make explicit.

3.4 Emerging Non-Volatile Memory Technologies

The storage tier of the memory hierarchy has undergone substantial transformation over the past two decades, driven primarily by the displacement of hard disk drives (HDDs) by NAND flash-based solid-state drives (SSDs) in a growing range of applications. This transition has had significant energy

implications, and the next wave of storage technology innovation — encompassing phase-change memory (PCM), spin-transfer torque magnetic RAM (STT-MRAM), and resistive RAM (RRAM) — promises further changes whose energy implications deserve careful examination.

Conventional hard disk drives consume energy for two distinct purposes: spinning the magnetic disk platters at speeds of 5,400 to 15,000 revolutions per minute, and moving the read-write head assembly across the disk surface to access data. Even in idle state, a hard disk drive consumes several watts — energy that is dissipated continuously whether or not any useful work is being performed. NAND flash SSDs, by contrast, consume energy only during active read and write operations, dropping to a low-power idle state in between. For workloads characterised by irregular, intermittent access patterns — which describe the majority of enterprise storage applications — SSDs therefore offer substantially lower energy consumption than HDDs, in addition to their better known advantages of lower latency and higher reliability.

Among the emerging non-volatile memory technologies, phase-change memory (PCM) — commercially available in limited form as Intel Optane, though Intel's retreat from the consumer Optane market in 2022 illustrates the commercial challenges of novel memory technologies — is of particular interest for green computing because it offers a combination of characteristics unavailable in any existing technology: byte-addressability like DRAM, non-volatility like NAND flash, and access latency intermediate between the two. A PCM cell stores a bit by switching between the amorphous and crystalline phases of a chalcogenide material, a transition achieved by heating the material with a brief electrical pulse. The energy per write

operation is somewhat higher than for DRAM, but the non-volatility of PCM means that data need not be periodically refreshed — eliminating the refresh power that accounts for a significant fraction of DRAM's idle energy consumption. For workloads that maintain large, slowly-changing datasets in memory, PCM-based storage class memory could offer meaningful total energy savings compared to DRAM.

Table 3.2 Storage Technology Energy and Endurance Comparison (2024)

Technology	Access Energy (read)	Idle Power	Typical Endurance	Volatility	Best Application Domain
DRAM (DDR5)	~8–12 pJ/bit	~0.5–2 W (active refresh)	~10^{15} cycles (effectively unlimited)	Volatile (requires power)	Main memory; working data sets
NAND Flash SSD (3D TLC)	~1–2 pJ/bit	~0.05–0.5 W idle	~300–1,000 P/E cycles	Non-volatile	Storage; OS; databases; warm data
NAND Flash SSD (3D SLC)	~1 pJ/bit	~0.05–0.3 W idle	~50,000–100,000 P/E cycles	Non-volatile	Write-intensive enterprise workloads
Phase-Change Memory (PCM)	~50–100 pJ/bit (write)	~0.01 W idle	~10^{7}–10^{8} cycles	Non-volatile	Storage-class memory; persistent data structures
STT-MRAM	~0.1–1 pJ/bit	~0 W (no refresh)	~10^{12} cycles	Non-volatile	On-chip cache replacement; IoT non-volatile RAM
Hard Disk Drive (7,200 RPM)	~100–500 pJ/bit	~4–8 W (always spinning)	~10^{6} load cycles	Non-volatile	Cold archival storage; large sequential workloads

P/E = Programme/Erase cycle. Idle power figures for SSDs and PCM assume the device is powered but not actively accessed. HDD idle power is substantially higher because the platters must be kept spinning. Sources: JEDEC standards documentation; Mutlu et al. (2019); manufacturer technical briefs.

3.5 Chiplet Architectures and Disaggregated Compute

The most recent significant development in hardware architecture that bears upon energy efficiency is the emergence of chiplet-based design — the assembly of a complete computing package from multiple smaller, individually manufactured semiconductor dies rather than from a single monolithic die. This approach, pioneered commercially by AMD with its Zen architecture processors and now adopted widely across the industry, has significant implications for both manufacturing economics and energy efficiency.

The energy efficiency dimension of chiplet design arises from several sources. First, smaller individual dies achieve higher manufacturing yields — fewer dies are wasted due to manufacturing defects — which reduces the energy cost per functional die produced. Second, chiplet architectures enable different parts of a processor to be manufactured in different process technologies optimised for their specific function: the compute cores might be fabricated at the most advanced and energy-efficient process node, while the I/O interface logic — which does not benefit as strongly from advanced scaling — is manufactured at a more mature, less expensive process node. This combination of ‘best process for each function’ can yield a more energy-efficient system than a monolithic die that must compromise all functions to a single process. Third, chiplet designs enable modular reuse of proven design blocks, reducing the development time and energy cost of bringing new products to market.

For the Indian semiconductor design community, chiplet architectures represent an opportunity that merits deliberate attention in the national semiconductor strategy. India's current strength lies primarily in chip design and verification rather than in advanced fabrication, and chiplet architectures reduce the dependence of a successful chip design on access to the most advanced fabrication nodes. A well-designed chiplet for a specific application — network security processing, agricultural IoT sensing, or railway communications management, to name three areas of obvious national relevance — can be assembled into a functional system package alongside commercially available chiplets from other manufacturers, without requiring access to 3-nanometre fabrication technology that India does not currently possess and cannot quickly acquire.

Green Data Centres

From Power Delivery to Cooling Innovation

Chapter 4

Green Data Centres — From Power Delivery to Cooling Innovation

In the summer of 2023, the operators of a large co-location data centre in Hyderabad faced a situation that their facility's original designers had not anticipated with sufficient seriousness. The monsoon, which was expected to moderate the city's temperatures and provide some relief to the cooling infrastructure, arrived late. For three weeks in June, ambient temperatures exceeded 42°C — eight degrees above the design assumptions embedded in the facility's air-side economiser system, a cooling mechanism that uses outdoor air directly when it is cool enough to do so. The economisers sat idle, the mechanical cooling systems ran continuously at full capacity, and the facility's Power Usage Effectiveness — the ratio of total facility power to IT equipment power — rose from its rated 1.4 to above 1.9 for those three weeks. The additional electricity cost, at the tariffs applicable to large commercial consumers in Telangana, was substantial. More fundamentally, the incident revealed a design vulnerability that climate projections suggest will become increasingly relevant across the Indian data centre sector in the decades ahead.

The management of energy in a data centre is an engineering discipline of considerable depth and complexity. It is the purpose of the present chapter to examine that discipline systematically, from the metrics used to characterise performance, through the engineering of power delivery and cooling systems, to the emerging frontier of waste heat recovery and renewable energy integration.

4.1 Metrics and Their Meanings: PUE, WUE, and Beyond

Any serious discussion of data centre energy efficiency must begin with the metrics through which that efficiency is measured and reported, for the reason that measurement frames perception, and perception drives action. The Power Usage Effectiveness (PUE) metric — defined as the ratio of total facility power consumption to the power consumed by IT equipment alone — has been the dominant metric of data centre energy efficiency since its introduction by The Green Grid consortium in 2007. Its widespread adoption represents a genuine achievement in standardising the language of data centre efficiency, enabling comparisons across facilities and creating a basis for industry benchmarking that did not previously exist.

PUE is computed as the ratio of total facility power to IT equipment power. A PUE of 1.0 would represent a theoretically perfect facility in which all energy consumed went directly into IT equipment and none was consumed by supporting infrastructure such as cooling, lighting, power conditioning, or management systems. In practice, the lowest PUEs achieved by leading hyperscale operators — Google has reported annual average PUEs below 1.12 for several of its most advanced facilities — require sophisticated engineering of every component of the facility's infrastructure. The global average PUE, as reported by the Uptime Institute's annual survey, remained approximately 1.55 to 1.58 in 2023, reflecting the large installed base of older, less well-optimised facilities that dominate the global fleet despite the efficiency of the newest hyperscale campuses.

It would be a mistake, however, to regard PUE as a complete or unambiguous measure of data centre sustainability. Several important limitations of the metric merit explicit

acknowledgement. First, PUE captures only the energy consumed within the facility boundary; it says nothing about the carbon intensity of the electricity used. A facility with a PUE of 1.1 powered entirely by coal-fired electricity has a larger carbon footprint than a facility with a PUE of 1.5 powered entirely by renewable hydroelectricity. Second, PUE is not normalised by the useful work performed by the IT equipment. A facility whose servers are poorly utilised — running at 10 per cent average CPU utilisation, which is not uncommon in older enterprise data centres — may achieve an impressive PUE while delivering relatively little computation per unit of energy. Third, PUE is susceptible to measurement gaming: the choice of measurement boundary, the averaging period, and the treatment of on-site generation can all be adjusted in ways that improve the reported figure without improving actual efficiency. These limitations are not arguments against the use of PUE, but they are arguments for using it alongside complementary metrics.

The Water Usage Effectiveness (WUE) metric, also developed by The Green Grid, captures the volume of water consumed by the facility per kilowatt-hour of IT equipment energy. As was noted in Chapter 1, this metric is of growing importance in the Indian context, where several major data centre markets face freshwater stress. The Carbon Usage Effectiveness (CUE) metric — the mass of CO_2 equivalent emitted per kilowatt-hour of IT equipment energy — captures the carbon intensity dimension that PUE ignores. The Energy Reuse Effectiveness (ERE) metric adjusts for waste heat recovery: a facility that reuses its waste heat to provide heating to adjacent buildings or industrial processes effectively reduces its net energy consumption, and ERE captures this benefit by subtracting reused energy from the numerator.

Table 4.1 Data Centre Efficiency Metrics — What They Measure and What They Miss

Metric	Definition	Scale	What It Captures	What It Misses	Status in India
PUE	Total facility power ÷ IT equipment power	1.0 (ideal) to ~3.0 (poor)	Overhead energy of cooling, power conditioning, lighting	Carbon intensity; utilisation; water consumption	Recommended by BEE; not yet mandatory reporting
WUE	Litres of water consumed ÷ IT equipment kWh	0.0 (ideal) to ~5.0 (poor)	Direct water consumption for cooling	Indirect water in grid electricity generation	Not yet widely reported by Indian operators
CUE	$kgCO_2e$ emitted ÷ IT equipment kWh	0.0 (ideal); varies by grid	Carbon intensity of facility energy supply	Embodied carbon of hardware; supply chain emissions	Voluntary; limited disclosure in India
ERE	(Total energy − reused energy) ÷ IT energy	0.0 (ideal) to ~1.0 or above	Benefit of waste heat recovery programmes	Requires third-party verification of reuse claims	Rarely reported; waste heat reuse uncommon in India
SCI	Carbon emissions per functional unit of software	Application-specific	Software-level carbon efficiency	Facility infrastructure; hardware lifecycle	Emerging metric; Green Software Foundation framework
Average Server Utilisation	Mean CPU/GPU utilisation across fleet (%)	0–100%	Efficiency of workload consolidation	Not captured in any standard facility metric	Not publicly reported; significant improvement opportunity

Sources: The Green Grid (2022); Green Software Foundation (2024); Bureau of Energy Efficiency, India (2022). BEE = Bureau of Energy Efficiency. The SCI metric is described fully in Chapter 6. The absence of mandatory WUE and CUE reporting in India represents a significant gap in the accountability framework for data centre sustainability.

4.2 Power Delivery: From the Grid to the Chip

The path that electrical energy travels from the utility grid to the transistors in a server rack is not a lossless one. At each stage of conversion and distribution within a data centre — from high-voltage alternating current at the facility connection point, through step-down transformers, uninterruptible power supplies, power distribution units, server power supplies, and voltage regulators on the server motherboard — a fraction of the energy is lost as heat. The cumulative efficiency of this power conversion chain is a significant contributor to PUE and, consequently, to the total energy consumed by the facility.

Traditional data centres used a power chain in which high-voltage AC from the utility was stepped down, rectified to DC, inverted back to AC for distribution through the facility, then rectified again to DC for use by server components. Each conversion step incurred losses, and the cumulative efficiency of the chain was typically in the range of 75 to 85 per cent — meaning that 15 to 25 per cent of the energy entering the facility was lost before reaching the IT equipment it was meant to power. Modern high-voltage direct current (HVDC) distribution architectures eliminate several conversion steps by distributing DC power at 380 or 480 volts directly to the server racks, where it is stepped down to the voltages required by individual components. HVDC distribution can achieve power chain efficiencies of 95 per cent or above, representing a substantial reduction in conversion losses.

Uninterruptible power supplies (UPS) represent a particular area of energy loss in traditional data centre designs. A UPS system that maintains a battery bank in a constant state of charge readiness, accepting AC power from the grid and converting it to DC for battery charging before converting it back to AC for the connected loads, incurs double-conversion losses of 2 to 10 per cent even during normal, non-outage operation. Online UPS systems — which run all power through the battery conversion path continuously — are significantly less efficient than offline or line-interactive UPS systems. The most energy-efficient approach in modern hyperscale designs is to eliminate the traditional UPS entirely, replacing it with distributed battery systems at the rack or server level that engage only during actual power interruptions, eliminating the continuous conversion losses.

4.3 Cooling Architectures: The Engineering Battle Against Heat

The cooling of data centres is, at its core, a thermodynamic engineering challenge: the removal of heat from a volume in which it is generated at high density, and its transfer to the external environment, with the minimum expenditure of additional energy. The three principal cooling architectures in current use — air cooling, direct liquid cooling, and immersion cooling — represent different engineering approaches to this challenge, each with distinct advantages, limitations, and energy implications.

Air cooling — the movement of cooled air through the server hall by computer room air conditioning (CRAC) or computer room air handler (CRAH) units, and the exhaustion of heated air through a hot aisle containment system — remains the most prevalent cooling approach globally, and the one with which most data centre operators have the greatest operational experience. Its energy

efficiency depends critically on the temperature differential between the supply air and the maximum allowable server inlet temperature: the higher the server can tolerate its inlet air temperature, the warmer the air that must be supplied, and the lower the energy required for cooling. ASHRAE's Thermal Guidelines for Data Processing Environments, which define allowable inlet temperature ranges for IT equipment, have been progressively revised upward as server manufacturers have qualified their equipment for higher operating temperatures, enabling operators to raise cooling set points and improve efficiency. The adoption of free-air cooling — drawing outdoor air directly into the facility for cooling when ambient temperatures are favourable — is an important extension of this principle, and is the basis for the extremely low PUEs achieved by hyperscale facilities in cold climates such as the Facebook data centres in Lulea, Sweden, and the Google facility in Hamina, Finland.

Direct liquid cooling (DLC) places cooling liquid — typically water or a water-glycol mixture — in direct thermal contact with the heat-generating components of the server, either through cold plates attached to processors and memory modules, or through rear-door heat exchangers that capture the heat of the server's exhaust air stream before it re-enters the room. DLC is significantly more effective than air cooling at removing heat from high-power components: the thermal conductivity and heat capacity of water are approximately 25 times those of air, allowing much greater heat removal per unit of volume and enabling server designs with much higher power densities than air cooling can accommodate. For the GPU-dense server configurations required for large-scale AI training — in which power densities of 40 to 100 kilowatts per rack are not uncommon — direct liquid cooling is increasingly a technical necessity rather than an efficiency option.

Immersion cooling represents the most aggressive approach to liquid cooling, submerging server hardware entirely in a thermally conductive but electrically insulating dielectric fluid — either a mineral oil or a synthetic engineered fluid such as the 3M Novec family. The fluid absorbs heat directly from all surfaces of the submerged hardware, achieving heat transfer efficiencies far superior to air cooling or even DLC cold plates, and enabling IT equipment to be operated at extreme power densities. Single-phase immersion systems, in which the fluid remains liquid, require external heat exchangers to reject the absorbed heat. Two-phase immersion systems exploit the latent heat of vaporisation of low-boiling-point fluids: the fluid boils on the hot hardware surfaces, the vapour rises and condenses on a cooled surface above, and the condensate returns by gravity — a passively efficient heat transfer mechanism that requires no pumping energy. It is pertinent to note that immersion cooling, despite its impressive efficiency credentials, has not yet achieved widespread adoption in mainstream data centre operations, primarily because it requires specialised equipment, complicates hardware servicing, and demands careful management of the fluid.

Table 4.2 Data Centre Cooling Architecture Comparison

Cooling Approach	Typical PUE Contribution	Max Rack Power Density	Water Consumption	Capital Cost (relative)	Suitable For
Air cooling (traditional CRAC)	1.4–2.0	~5–10 kW/rack	High (evaporative towers)	Low	Legacy facilities; low-density workloads; cost-sensitive deployments

Air cooling + free-air economisation	1.1–1.4	~10–15 kW/rack	Medium–Low (climate-dependent)	Low–Medium	Facilities in cool climates; newer enterprise deployments
Direct Liquid Cooling (cold plates)	1.05–1.2	~30–60 kW/rack	Low–Medium (closed loop possible)	Medium	High-performance computing; AI training clusters; modern hyperscale
Rear-door heat exchanger	1.1–1.3	~15–25 kW/rack	Medium (requires chilled water)	Low–Medium	Retrofit of existing air-cooled facilities with high-density rows
Single-phase immersion cooling	1.02–1.1	~100+ kW/rack	Very Low (no evaporation)	High	Extreme density AI GPU clusters; edge deployments in harsh environments
Two-phase immersion cooling	1.01–1.05	~150+ kW/rack	Near-zero direct consumption	Very High	Research; specialised HPC; emerging for AI at maximum density

PUE contribution refers to the cooling system's contribution to total facility PUE; IT power conversion losses are additional. Capital cost is relative and highly site-dependent. Sources: Uptime Institute (2024); ASHRAE TC 9.9 (2021); Lawrence Berkeley National Laboratory Data Center Research Programme (2023).

4.4 Renewable Energy and the 24/7 Carbon-Free Energy Challenge

The procurement of renewable electricity — through power purchase agreements (PPAs), renewable energy certificates (RECs), or on-site generation — has become a standard element of the sustainability strategy of major data centre operators. Google has maintained that it matches its annual electricity consumption with renewable energy purchases since 2017; Microsoft has committed to powering its facilities with 100 per cent renewable energy; Amazon Web Services has become one of the largest corporate purchasers of renewable power in the world. These commitments are genuine and, within their methodological scope, meaningful. They are also, however, subject to an important limitation that the more discerning sustainability analysts have begun to articulate: the difference between annual matching and 24/7 carbon-free energy.

Annual matching — the most common approach — involves purchasing renewable energy certificates equivalent in total to the facility's annual electricity consumption, without regard to when during the year that energy was generated. A data centre that operates around the clock in a region with limited renewable generation might match its annual consumption by purchasing RECs generated during daylight hours by solar farms in a different region, or during windy months by wind farms in yet another. The electricity actually consumed by the facility at three in the morning in December may be generated entirely by coal plants; the RECs purchased offset this on paper but do not change the physical reality of what powers the facility at that moment.

The 24/7 carbon-free energy (CFE) concept — championed by Google and formalised in a framework developed in collaboration with UN Energy — sets a more demanding standard:

the goal is to match the facility's energy consumption with carbon-free generation on an hourly basis, in the same grid region. Achieving 24/7 CFE requires a combination of technologies — solar and wind for daytime generation, battery storage for evening and overnight supply, and firm power from geothermal, nuclear, or hydroelectric sources for continuous baseload — that is considerably more expensive and technically demanding than annual matching. As of 2024, Google had achieved 24/7 CFE at a limited number of its facilities and had committed to achieving it globally by 2030; the broader industry remains at various stages of progress toward this more demanding standard.

For Indian data centre operators, the renewable energy landscape presents both opportunities and challenges that are specific to the national context. India's solar energy capacity has grown remarkably, reaching over 80 gigawatts of installed capacity by early 2024, and the cost of utility-scale solar power in India — which has fallen to among the lowest in the world, with recent auction prices below Rs 2.50 per kilowatt-hour — makes solar PPAs an economically attractive option for large data centre operators. The challenges are the intermittency of solar generation and the limited development of grid-scale battery storage at competitive cost in India. Data centres in Tamil Nadu, Rajasthan, and Gujarat — states with strong solar resources and progressive renewable energy policies — are better positioned to pursue meaningful renewable matching than those in states with less favourable conditions. The Ministry of New and Renewable Energy's recent push to develop offshore wind capacity, and the development of pumped hydro storage projects in the Western Ghats and Himalayas, will improve the long-term prospects for 24/7 CFE in India, but these developments are measured in decades rather than years.

4.5 Waste Heat Recovery: From Liability to Asset

The heat removed from a data centre by its cooling systems is, in one sense, a waste product: energy that was purchased for computation, used to power electronic components that converted it mostly to heat, and then expended additional energy to remove that heat from the building. In another sense, it is a resource: a large, continuous, and predictable source of thermal energy that, with appropriate engineering, can be put to useful work rather than simply rejected to the atmosphere.

Waste heat recovery from data centres is not a new concept, but its practical implementation remains relatively uncommon, primarily because it requires a suitable demand for low-grade heat in close proximity to the data centre. The heat rejected by most air-cooled data centre cooling systems is available at temperatures of 25 to 40 degrees Celsius — low-grade heat that is unsuitable for most industrial processes, which require higher temperatures, but that is well-suited to space heating, domestic hot water, and greenhouse heating applications. The most successful examples of data centre waste heat reuse in practice are located in Scandinavia, where district heating networks provide a ready infrastructure for distributing the recovered heat to residential and commercial buildings. The data centres of Fortum in Helsinki and those operated by various operators in Stockholm supply heat to district heating networks serving hundreds of thousands of residents.

In the Indian context, direct analogues to district heating are uncommon, given the climate. However, the application of data centre waste heat to industrial drying processes, aquaculture operations, or greenhouse cultivation represents potentially viable avenues for waste heat recovery in appropriate locations. The higher coolant temperatures available from immersion and liquid-cooled systems — which can reject heat at 50 to 60 degrees Celsius

or above, compared to the 25 to 35 degrees available from air-cooled systems — expand the range of potential end uses and make absorption cooling an option: the waste heat can be used to drive an absorption chiller, which produces cooling for an adjacent facility, effectively converting waste heat from the data centre into cooling for a neighbour.

Green Networking

The Energy Cost of Moving Bits

Chapter 5

Green Networking — The Energy Cost of Moving Bits

It is an instructive exercise in quantitative imagination to consider what happens, in physical terms, when a resident of Chennai sends a video message to a relative in Vancouver. The video data, encoded on the sender's smartphone, travels via a 4G or 5G radio link to the nearest base station — one of several hundred thousand base stations operated by India's telecommunications companies across the country. From there, it passes through the optical fibre backhaul network of the operator, across the undersea cable systems that connect India to North America — likely the SEA-ME-WE 6 or AAG systems, running along the ocean floor at depths of several kilometres — through the Canadian network infrastructure, and finally via a local radio or broadband link to the recipient's device. The total distance traversed may be twenty thousand kilometres. The energy consumed in this traversal, though small for a single message, is the aggregate of thousands of individual network elements, each of which has been designed with varying degrees of attention to energy efficiency, and each of which contributes to a total that, at the scale of billions of such messages daily, is very considerable indeed.

The energy consumption of the global telecommunications network — access networks, transmission networks, core networks, and the equipment that connects them — is estimated to be in the range of 200 to 400 terawatt-hours annually, comparable in magnitude to the energy consumption of the world's data centres. Yet this energy receives proportionally less attention in sustainability discussions, perhaps because it is more diffuse —

distributed across millions of pieces of equipment in thousands of locations — and because its growth has been more successfully masked by efficiency improvements to date. The present chapter examines the energy landscape of telecommunications networking, from the physics of radio transmission to the protocols of the global Internet, with particular attention to the implications of India's extraordinary ongoing investment in network infrastructure.

5.1 Energy Models for Telecommunications Networks

Before the energy consumption of a network can be reduced, it must be understood — which requires a model of how energy is consumed at each level of the network hierarchy. The telecommunications network can, for analytical purposes, be decomposed into three principal tiers, each with a distinct energy profile.

The access network — the portion of the network that connects end users to the telecommunications infrastructure — is responsible for a disproportionate share of total network energy consumption. This is because the access network is inherently distributed: it must deploy equipment close to every user, including users in low-density areas where the cost and energy of a base station or local exchange is amortised over very few connections. Mobile base stations — the radio transceivers that provide wireless connectivity to smartphones and other mobile devices — are the dominant component of access network energy consumption. A typical 4G base station consumes between 500 watts and 2 kilowatts of electrical power, of which a significant fraction is consumed in the radio frequency power amplifiers that transmit the radio signal. Since base stations must remain powered continuously regardless of whether they are actively serving users

— to provide coverage and respond to devices that wish to connect — the idle power of a base station that is lightly loaded at three in the morning is nearly the same as its peak power during the morning rush hour.

The metro network — which aggregates traffic from multiple access network nodes and carries it to the core — typically consists of optical fibre rings connecting switching and routing nodes within a metropolitan area. The energy consumption of the metro network is dominated by the optical amplifiers, transponders, and switching equipment at these nodes. Metro network equipment is somewhat more amenable to energy proportionality — the ability to reduce power consumption when traffic is light — than access equipment, though the improvement that is practically achievable remains limited by the overhead of control plane functions that must operate continuously.

The core network — the high-capacity long-distance transmission infrastructure that interconnects metropolitan areas, countries, and continents — is characterised by very high traffic volumes, long transmission distances, and sophisticated optical systems that require substantial infrastructure and energy investment. The energy per bit transmitted across the core network has fallen dramatically over the past two decades as optical amplification technology and dense wavelength-division multiplexing have improved, and the core network is now the most energy-efficient tier of the network on a per-bit basis, though its absolute energy consumption remains significant by virtue of the enormous volumes of traffic it carries.

Table 5.1 Network Energy Intensity by Segment — Selected Estimates (2022–2024)

Network Segment	Energy Intensity (kWh/GB)	Absolute Consumption (TWh, global)	Dominant Energy Driver	Energy Proportionality
Mobile access (4G LTE)	~0.1–0.5 kWh/GB	~60–100	Base station RF amplifiers + site cooling	Poor — near-constant power regardless of load
Mobile access (5G NR, sub-6 GHz)	~0.05–0.3 kWh/GB at peak load	Growing rapidly	Massive MIMO antenna arrays; increased site density	Moderate — sleep modes improving
Fixed broadband access (FTTH/GPON)	~0.01–0.05 kWh/GB	~20–40	Optical line terminals; CPE	Moderate — some load-dependent scaling
Metro aggregation network	~0.005–0.02 kWh/GB	~15–30	Optical transponders; routers; switches	Moderate — traffic shaping aids efficiency
National/long-haul fibre backbone	~0.001–0.005 kWh/GB	~20–40	Optical amplifiers; DWDM transponders	Good — energy scales with active wavelengths
Submarine cable systems	~0.001–0.003 kWh/GB	~5–10	Optical repeaters; cable ship operations	Good — fixed plant; improving with new cable generations
IP core routers and Internet exchanges	~0.002–0.01 kWh/GB	~15–30	Line card forwarding ASICs; cooling	Moderate — capacity overprovisioned for resilience

Energy intensity figures are highly variable across operators, equipment generations, and traffic conditions; values represent credible midpoints of published estimates. Sources: Coroama and Hilty (2014); Andrae (2020); International Telecommunication Union (2023); Malmodin and Lundén (2018).

5.2 The 5G Energy Equation: More Capability, More Consumption

The worldwide deployment of fifth-generation (5G) mobile networks, which began in earnest between 2019 and 2022 and is proceeding rapidly in India with the rollout of Jio and Airtel's 5G networks, offers substantially enhanced mobile connectivity — higher peak data rates, lower latency, and greater capacity for simultaneous connections — at a significant energy cost premium over the 4G networks it is supplementing and gradually replacing. Understanding the sources of this energy cost premium, and the measures available to moderate it, is of particular importance for India, which is simultaneously committed to aggressive network modernisation and to its energy and climate NDC targets.

The energy cost of 5G networks arises from two principal architectural changes relative to 4G. The first is the use of massive Multiple Input Multiple Output (massive MIMO) antenna systems, which deploy dozens to hundreds of active antenna elements per base station, each contributing to a beamformed radio signal that can be directed toward individual users rather than broadcast omnidirectionally. Massive MIMO dramatically improves spectral efficiency — the amount of data transmitted per unit of radio spectrum — but the large number of active antenna chains requires substantial computational and power amplification hardware that consumes considerably more energy than a conventional 4G antenna. A typical massive MIMO 5G base station consuming 2 to 5 kilowatts represents a two-to-three-fold increase in site power consumption relative to a comparable 4G installation.

The second architectural change is densification: the deployment of a much denser network of 5G base stations, particularly small cells and microcells operating in the millimetre-wave spectrum bands, to achieve the high data rates and low latency that 5G promises in dense urban environments. While each small cell consumes less power than a macro base station, the large number of additional sites required to achieve coverage adds significantly to total network energy consumption. Airtel has publicly stated that its early 5G deployments showed energy consumption two to three times that of equivalent 4G sites; the company has since invested substantially in energy efficiency measures — including AI-based network management and renewable energy for base station sites — to moderate this increase.

It is pertinent to note, however, that the energy cost of 5G networks, while real, should be assessed in the context of the capability it delivers. If the higher data throughput enabled by 5G allows more tasks to be completed in less time — video calls completed more quickly, large file transfers accomplished in seconds rather than minutes — and if this results in reduced overall time-on-air for the wireless network, the net energy impact per unit of useful communication may be more favourable than a comparison of site power consumption alone would suggest. The energy efficiency metric that matters for green networking is not power per site but energy per bit successfully delivered — and on this metric, 5G's higher spectral efficiency can result in improvements over 4G at high traffic loads, even if its absolute power consumption per site is higher.

Table 5.2 5G versus 4G LTE — Energy Profile Comparison

Parameter	4G LTE (Typical Macro Cell)	5G NR (Sub-6 GHz Macro)	5G NR (mmWave Small Cell)	Key Implication
Site power consumption	~500–1,500 W	~2,000–5,000 W	~100–500 W per small cell	Higher absolute power per macro site; many more small cell sites needed
Antenna configuration	2×2 or 4×4 MIMO (4–8 ports)	Massive MIMO (32–192 active ports)	Phased array (16–64 elements)	Massive MIMO greatly increases computation and RF hardware energy
Spectral efficiency (typical load)	~3–5 bits/s/Hz	~8–15 bits/s/Hz	~15–30 bits/s/Hz	Higher spectral efficiency means more bits per joule at high load
Energy per bit (at peak load)	~0.3–1.0 μJ/bit	~0.1–0.5 μJ/bit	~0.05–0.2 μJ/bit	5G can be more efficient per bit than 4G at high utilisation
Energy per bit (at 10% load)	~3–10 μJ/bit	~5–20 μJ/bit (worse)	~0.5–2 μJ/bit	5G less efficient than 4G at low utilisation — sleep modes critical
Sleep mode capability	Limited (CRS transmission required)	Enhanced NR sleep modes (PDCCH gating)	Better — small cells can power-cycle rapidly	NR sleep modes partially offset idle energy disadvantage
Renewable energy integration	Moderate — large sites with space for solar	Moderate–Good — same site footprint	Good — small cells suit rooftop solar	5G small cells well-suited for distributed renewable powering

μJ = microjoules. Energy per bit values are workload and load-dependent; peak load figures assume near-capacity utilisation. NR = New Radio (5G standard). PDCCH = Physical Downlink Control Channel. Sources: Fehske et al. (2011); Ericsson Technology Review (2022); TRAI Consultation Paper on Energy Efficiency in Telecom (2023); Airtel and Jio sustainability disclosures.

5.3 Software-Defined Networking and Energy-Aware Traffic Engineering

The traditional telecommunications network was built on purpose-specific hardware — routers, switches, amplifiers, and transponders each designed for a single function and operating under software that was tightly coupled to the hardware platform. This architecture, while reliable and well-understood, offers limited flexibility for energy optimisation: the network is dimensioned for peak traffic load, and the underutilisation of that capacity during off-peak periods represents a persistent source of energy waste. Software-Defined Networking (SDN) and Network Functions Virtualisation (NFV) are architectural approaches that address this limitation by decoupling the control logic of network functions from the physical hardware on which they run, enabling more flexible and energy-responsive management of network resources.

In an SDN-enabled network, a centralised controller maintains a global view of network topology, traffic loads, and equipment energy states, and can direct traffic flows along paths that minimise energy consumption as well as — or instead of — paths that minimise delay or maximise throughput. During periods of low traffic, the controller can consolidate all active traffic flows onto a subset of active links and nodes, powering down or putting to sleep the unused equipment. This is the network-level analogue of the processor sleep states described in Chapter 2: equipment that is not needed can be placed in a low-power state, eliminating its

idle energy consumption. The practical challenge is that the network's resilience and latency requirements constrain how aggressively this consolidation can be pursued — some equipment must remain active to ensure that traffic can be re-routed quickly in response to failures.

India's Department of Telecommunications has included energy efficiency requirements in the framework for 5G spectrum assignment, and the Telecom Regulatory Authority of India (TRAI) has published consultation papers on energy efficiency standards for telecommunications infrastructure. These are encouraging developments, though the translation from consultation papers to enforceable standards has historically been slow. The more immediate source of energy efficiency progress in Indian telecoms has been economic rather than regulatory: the very high energy costs faced by telecom operators in India, and the substantial share of those costs attributable to diesel generator operation at base stations in areas with unreliable grid power, have created strong commercial incentives for investment in solar-powered base stations and energy-efficient equipment. Bharti Airtel's Project Nxtra and Reliance Jio's initiatives to solarise their base station networks are commercially motivated as much as environmentally motivated, and the alignment of commercial and environmental incentives in this domain is a genuinely positive feature of the Indian telecoms landscape.

5.4 Content Delivery Networks and the Geography of Energy

A significant proportion of Internet traffic — video streaming, software updates, web content, and cloud application data — is delivered not from origin servers located in distant data centres but from edge caches operated by Content Delivery Networks (CDNs)

positioned close to users in regional or metropolitan network facilities. CDN caching reduces the distance that popular content must travel across the network to reach users, reducing both network latency and network energy consumption. The energy savings arise from two sources: the reduction in long-haul transmission energy for content that is served from a nearby cache rather than a distant origin, and the reduction in the load on origin data centres that do not need to serve every request individually.

The energy implications of CDN deployment are not, however, uniformly positive. CDN edge nodes consume their own energy, and the question of whether the energy saved in transmission exceeds the energy consumed by the edge infrastructure depends on the cache hit rate — the proportion of requests that can be served from the local cache rather than requiring a fetch from the origin — and on the efficiency of the edge node hardware relative to the origin data centre. For content with very high popularity — live sports broadcasts, new software release downloads, widely accessed web pages — cache hit rates are high and CDN deployment is clearly energy-beneficial. For long-tail content with low popularity, cache hit rates are lower and the energy calculus is less clear.

For India, the development of domestic CDN infrastructure is a matter of both economic and energy significance. Historically, a substantial proportion of the Internet content consumed by Indian users was served from CDN edge nodes located outside India — in Singapore, Hong Kong, or London — because the Indian Internet infrastructure, while growing rapidly, has not always provided cost-effective facilities for CDN deployment at scale. The consequence is that video streams and web content served to users in Mumbai sometimes travel thousands of kilometres to a distant edge node before returning to India. The rapid development of

domestic data centre capacity, and the improvement of peering arrangements at Internet exchange points such as the Mumbai Internet Exchange (MIX) and the Chennai Internet Exchange (CIX), is gradually improving this situation, reducing both the latency experienced by Indian users and the network energy consumed in serving them.

5.5 The Emerging Energy Challenge of India's Digital Infrastructure

The telecommunications sector in India is, by any measure, undergoing a transformation of historical significance. The number of broadband subscribers has grown from approximately 280 million in 2019 to over 900 million in 2024, a more than threefold increase in five years. The per-capita monthly data consumption of Indian mobile users, at over 20 gigabytes per month as of 2023, now exceeds that of users in many developed economies. The deployment of 5G networks by Reliance Jio and Bharti Airtel, which began in earnest in 2022, will bring high-capacity connectivity to hundreds of millions of additional users over the next five years. The government's BharatNet programme is extending broadband connectivity to the 600,000-plus gram panchayats of rural India, many of which have never had reliable connectivity before.

This expansion of digital access is, it must be emphasised at the outset, a profoundly positive development from the perspective of human welfare, economic opportunity, and social equity. The energy cost of universal digital access, properly assessed, is a cost worth bearing. The question that the green networking discipline poses is not whether to expand digital access, but how to expand it in the most energy-efficient manner possible — which combination of technologies, architectures, and operational

practices will deliver universal connectivity while minimising the energy and carbon cost per connected user.

Several principles emerge from the analysis in the present chapter that are directly applicable to the Indian context. First, the solarisation of base stations — particularly in areas where grid connectivity is unreliable and diesel generation is otherwise the fallback — represents one of the highest-value energy interventions available to Indian telecom operators, both economically and environmentally. Second, the design of BharatNet's rural infrastructure should incorporate the latest generation of energy-efficient equipment and, wherever possible, solar or other renewable energy sources, rather than replicating the energy profile of earlier generations of urban network infrastructure. Third, the development of domestic CDN infrastructure and the improvement of Internet exchange peering within India will reduce the energy cost of the content delivery that constitutes the majority of Indian mobile data traffic. Fourth, the adoption of SDN-based energy management in India's growing metropolitan fibre networks offers a practical pathway to significant efficiency improvements without requiring replacement of the physical infrastructure.

Green Software Engineering

Writing Code That Respects the Planet

Chapter 6

Green Software Engineering — Writing Code That Respects the Planet

In the year 2020, a team of researchers at the University of Minho in Portugal published a study that deserves to be read by every student of computer science and software engineering. The study measured the energy consumption of identical computational tasks implemented in twenty-seven different programming languages — sorting, binary tree manipulation, spectral norm calculation, and other benchmark operations — and found that the energy consumed varied by a factor of more than eighty between the most and least efficient languages. C, the language closest to the hardware, was the most energy-efficient. Perl, Ruby, and Python were among the least efficient. The differences were not trivial: a Python implementation of a task that a C programme completed with a given quantity of energy might require seventy-five times as much.

Those who teach programming will recognise an uncomfortable implication in this finding. For three decades, computer science education has been moving steadily away from low-level languages toward high-level ones — toward Python, toward JavaScript, toward interpreted and dynamically typed languages that reduce programming friction at the cost of computational efficiency. This trend has been, in many ways, a productive and necessary one. But it has been pursued, for the most part, without accounting for the energy cost of the abstraction it provides. The present chapter argues that this accounting must now be done, and that the discipline of green software engineering — designing, writing, and deploying code with explicit attention to its

energy consumption — is not a speciality topic for a minority of practitioners but a professional responsibility that belongs to every engineer who writes software that runs at scale.

6.1 Software Is Not Weightless: The Energy Cost of Abstraction

There is a conception of software that is widely held, implicitly if not explicitly, among practitioners and non-practitioners alike: that software is immaterial, a pure product of thought untethered from physical constraints. This conception has a certain philosophical charm, but it is, from the perspective of energy and sustainability, profoundly misleading. Every instruction executed by a processor consumes energy. Every byte moved between memory and processor consumes energy. Every network packet transmitted consumes energy. Software, which ultimately reduces to sequences of instructions, memory accesses, and network operations, is not immaterial at all; it is a specification for physical work, and the efficiency with which it is written determines how much physical work — and therefore energy — is required to accomplish any given task.

The energy cost of software arises from several sources that operate at different levels of the system stack, and it is pertinent to understand each of these sources in order to appreciate where the opportunities for green software engineering lie.

At the algorithmic level, the computational complexity of an algorithm — the number of operations required to solve a problem of a given size — is the most fundamental determinant of its energy consumption at scale. An algorithm that solves a problem in $O(n \log n)$ time will consume less energy, for large inputs, than one that solves the same problem in $O(n^2)$ time, because it performs fewer operations. This is a well-understood principle of algorithm

analysis that requires no revision for the energy context. What does require revision, however, is the common assumption that computational complexity is the only relevant dimension of algorithmic energy cost. As was established in Chapter 2, memory access patterns can dominate energy consumption for memory-bound algorithms, and an $O(n^2)$ algorithm with excellent cache locality may consume less energy than an $O(n \log n)$ algorithm with poor locality. Energy-aware algorithm design must therefore consider both computational complexity and memory access patterns simultaneously.

At the implementation level — the choices made about data structures, programming language, library selection, and code organisation — the energy consequences are substantial and often under-appreciated. The Pereira et al. (2017) study referenced in the opening vignette of the present chapter made this quantitative and public in a way that had not previously been done. Their finding that Python requires approximately seventy-five times as much energy as C to perform equivalent computational tasks reflects the overhead of interpretation, dynamic type checking, garbage collection, and the multiple layers of abstraction that high-level language runtimes impose. This does not mean that Python should be abandoned — its productivity advantages for data science, machine learning, and rapid prototyping are genuinely valuable, and for applications that are not computationally intensive, the energy overhead is inconsequential. What it does mean is that when Python is used for computationally intensive tasks at scale — processing large datasets, running simulation loops, serving high-volume API endpoints — the energy cost of that choice should be understood and, where it is significant, mitigated through appropriate techniques.

At the deployment and operations level, decisions about how software is run — the degree of resource provisioning, the utilisation of deployed capacity, the scheduling of workloads in time — have energy consequences that can equal or exceed the consequences of algorithmic and implementation choices. A well-written programme deployed on massively over-provisioned hardware, running at 5 per cent utilisation, consumes far more energy per unit of useful work than a less polished programme deployed on appropriately sized hardware at 70 per cent utilisation. Server utilisation rates in enterprise data centres — including those operated by Indian public sector organisations and private sector companies — have historically been very low, often in the range of 10 to 15 per cent, representing a substantial and largely addressable source of energy waste.

6.2 Programming Language Energy Efficiency: Evidence and Implications

The empirical investigation of programming language energy efficiency is a relatively recent subdiscipline of computer science, but it has accumulated sufficient evidence to support some reasonably firm conclusions. The study by Pereira and colleagues at the University of Minho, published in 2017 and subsequently extended in 2021, remains the most comprehensive publicly available analysis of this question, and its findings merit examination in some detail.

The study used the Computer Language Benchmarks Game — a curated set of computational tasks designed to measure the performance of programming language implementations — to measure energy consumption, execution time, and peak memory usage across twenty-seven languages. The energy measurements were performed using RAPL (Running Average Power Limit),

Intel's hardware energy measurement interface, which provides accurate, high-frequency measurements of processor and DRAM energy consumption. The findings, expressed as multiples of the most efficient language's energy consumption for each task, revealed a striking and consistent pattern: compiled, statically typed languages — C, C++, Rust, Ada — consumed the least energy, typically within a factor of two to three of one another. Just-in-time compiled languages — Java, C#, Go — occupied a middle tier, typically consuming five to ten times the energy of the compiled languages. Interpreted, dynamically typed languages — Python, Ruby, Perl — consumed the most energy, with some implementations exceeding seventy-five times the energy of the most efficient.

It is necessary to interpret these findings with some care before drawing prescriptive conclusions. The benchmark tasks used in the study are computational kernels — tight numerical loops, sorting algorithms, matrix operations — that favour compiled languages by design. For tasks that are I/O-bound rather than compute-bound — reading files, making network requests, waiting for database responses — the language's execution overhead is often irrelevant because the computation is not the bottleneck. For tasks involving developer time rather than machine time — rapid prototyping, exploratory data analysis, one-off scripting — the productivity advantages of high-level languages may represent an indirect energy saving by reducing the computation required to iterate toward a correct solution. The appropriate lesson from Pereira et al. is not that all software should be rewritten in C, but that language choice should be a deliberate decision that considers energy implications alongside developer productivity, maintainability, and ecosystem suitability.

For the Indian software industry — which employs approximately five million software professionals and produces software consumed globally across every domain — the aggregate energy implications of language and implementation choices are not negligible. India's software exports, valued at over USD 200 billion in the financial year 2023–2024 as reported by NASSCOM, represent an enormous quantity of deployed software running continuously on servers around the world. If that software were, on average, even 10 per cent more energy-efficient — through better algorithmic choices, more appropriate language selection for compute-intensive components, and higher server utilisation rates — the aggregate reduction in global data centre energy consumption would be measurable.

Table 6.1 Programming Language Energy Efficiency — Selected Results

Language	Relative Energy (C = 1.0)	Relative Time	Relative Memory	Type System	Best Used For
C	1.00 (reference)	1.00	1.00	Static, manual	Systems programming, embedded, performance-critical kernels
C++	1.34	1.56	1.34	Static, manual/RAII	Systems, game engines, HPC, high-performance applications
Rust	1.03	1.04	1.54	Static, ownership	Systems programming, WebAssembly, safety-critical embedded

Java	1.98	1.89	6.01	Static, JVM managed	Enterprise applications, Android, large-scale backend services
C# (.NET)	3.14	2.85	2.85	Static, CLR managed	Enterprise, Windows, game scripting (Unity), web APIs
Go	3.23	2.83	4.61	Static, GC managed	Cloud-native services, network tools, DevOps tooling
JavaScript (Node.js)	4.45	6.52	4.59	Dynamic, JIT	Web front-end, lightweight APIs, scripting
Python 3	75.88	71.90	2.80	Dynamic, interpreted	Data science, ML research, scripting, rapid prototyping
Ruby	69.91	59.34	6.72	Dynamic, interpreted	Web applications, scripting, rapid development
Perl	79.58	65.07	6.62	Dynamic, interpreted	Text processing, legacy scripting, sysadmin tools

Relative values expressed as multiples of C's energy, time, and memory consumption on the Computer Language Benchmarks Game tasks. These are computational kernel benchmarks; I/O-bound applications show smaller differences. Source: Pereira et al. (2017, 2021), University of Minho. Values represent geometric means across multiple benchmark tasks and should be treated as indicative orders of magnitude rather than precise predictions for specific applications.

6.3 Carbon-Aware Scheduling: Exploiting Temporal Variation in Grid Carbon Intensity

One of the most practically accessible — and most underutilised — opportunities for reducing the carbon emissions of computing is temporal shifting: the deliberate scheduling of flexible workloads to run at times when the carbon intensity of the electricity grid is lower. This strategy requires no change to the software itself, no hardware upgrade, and no reduction in the useful work performed; it requires only the willingness to accept some delay in the execution of non-urgent tasks, and the infrastructure to know when the grid is clean.

The carbon intensity of electricity grids varies substantially by time of day, day of the week, and season, primarily because the mix of generation sources varies continuously. Solar generation is concentrated during daylight hours; wind generation varies with weather patterns; thermal generation serves as backup when renewable generation is insufficient. In regions with substantial renewable capacity, the carbon intensity of the grid can vary by a factor of three to five between the cleanest and dirtiest hours of the day. A batch processing job that can tolerate being run at two in the morning rather than two in the afternoon — when solar generation is at its peak and grid carbon intensity is at its daily minimum in a solar-rich region — may emit substantially less carbon for exactly the same computational work.

The Software Carbon Intensity (SCI) specification, developed by the Green Software Foundation and published in 2023, provides a standardised framework for quantifying the carbon emissions of software per functional unit of work. The SCI score is computed as: SCI = (E × I) + M, where E is the energy consumed per functional unit, I is the marginal carbon intensity of the electricity grid at the time and location of execution, and M is the embodied

carbon of the hardware, amortised over its operational life. The inclusion of the grid carbon intensity factor I explicitly in the metric — rather than treating energy consumption as a proxy for carbon — makes the temporal and geographic variation in grid cleanliness directly relevant to software optimisation decisions. A software system that is aware of its SCI score can, in principle, schedule flexible work to minimise that score by exploiting low-I periods.

In the Indian context, the application of carbon-aware scheduling requires engagement with the complexity of India's regional electricity grids. The carbon intensity of electricity varies significantly across India's five regional grids — the Northern, Eastern, Western, Southern, and North-Eastern regional grids — reflecting the different generation mixes of the states they serve. The Southern Regional Grid, which serves Tamil Nadu, Karnataka, Andhra Pradesh, Telangana, and Kerala, has a substantially lower average carbon intensity than the Northern Regional Grid, which serves Punjab, Haryana, Uttar Pradesh, and Delhi, partly because of the higher share of renewables in the southern states. Data centre operators with the flexibility to place workloads across multiple geographic locations — cloud operators, in particular — can exploit this regional variation in addition to temporal variation.

6.4 Green Software Patterns: Practical Design Principles

Beyond language choice and scheduling strategy, a set of software design patterns can be identified that consistently produce more energy-efficient implementations than their alternatives. These patterns are not arcane or specialised; they are, for the most part, extensions of existing good software engineering practice that have been given an explicit energy rationale.

Lazy evaluation — the deferral of computation until its results are actually needed — reduces energy waste by avoiding computations whose results may never be used. In a system that generates a report containing fifty fields but whose users access an average of three fields per session, eagerly computing all fifty fields for every request wastes the energy of forty-seven computations per session. Lazy evaluation computes each field only when it is requested, reducing per-session energy consumption proportionally to the average fraction of fields actually accessed. This principle applies at every level of the system stack, from individual function calls to API response construction to database query planning.

Caching — the storage of computed results for reuse in response to repeated identical requests — is perhaps the most widely understood performance optimisation in software engineering, and it is simultaneously a powerful energy optimisation. A request served from a cache consumes a fraction of the energy of a request that requires a full computation: cache hits typically require only a memory access and a network response, while cache misses may require database queries, complex computations, and multiple network round trips. For applications with high request repetition rates — which describes the majority of web services, including the high-traffic applications of India's digital public infrastructure such as DigiLocker, UMANG, and the GSTN — appropriately designed caching layers can reduce per-request energy consumption by an order of magnitude or more.

Batching — the aggregation of multiple small operations into fewer large ones — reduces the overhead costs that attend every individual operation regardless of its size: network round-trip latency, database transaction overhead, function call overhead, and

the waking of processors from sleep states. A system that sends one hundred individual database write requests consumes substantially more energy than one that batches those writes into a single transaction, because the overhead cost is incurred once rather than one hundred times. Batching is a pattern that is particularly relevant to IoT and mobile applications, where battery-powered devices must manage energy carefully and where the energy cost of radio transmission — the dominant energy cost of mobile devices for many application patterns — is proportional to the number of transmissions rather than their size.

It is pertinent to note, in the context of the present discussion, that these patterns — lazy evaluation, caching, and batching — are not uniquely motivated by energy considerations. They are established software engineering best practices whose energy benefits are an additional reason to apply them. The discipline of green software engineering does not, in the main, require engineers to choose between good software and efficient software; it reveals that good software engineering practice and energy efficiency are, more often than not, aligned. The cases where they diverge — where the most energy-efficient approach requires sacrificing maintainability or developer productivity — are real but less common than one might suppose.

Table 6.2 Software Carbon Intensity (SCI) Metric — Components and Measurement

SCI Component	Symbol	Definition	Measurement Approach	Reduction Strategy
Energy per functional unit	E	kWh of electricity consumed to deliver one	Hardware energy measurement (RAPL, IPMI, PDU meters);	Algorithm optimisation; language selection; caching;

		functional unit of work	cloud provider billing APIs	batching; hardware right-sizing
Grid carbon intensity	I	gCO_2eq per kWh at time and location of execution	Electricity Maps API; WattTime API; regional grid operator data (POSOCO for India)	Temporal shifting to low-carbon hours; geographic placement on low-carbon grids
Embodied carbon (amortised)	M	gCO_2eq per functional unit from hardware manufacturing, divided by expected lifetime use	LCA databases (ecoinvent, GaBi); manufacturer EPD documents; Cloud Carbon Footprint tool	Hardware longevity; right-sizing to avoid over-provisioning; circular economy procurement
Functional unit	R	The unit of work the software delivers (e.g., per API call, per user session, per GB processed)	Application-level instrumentation; logging and monitoring infrastructure	Definition must be stable for SCI to be comparable over time
SCI score	SCI	(E × I) + M, expressed per functional unit R	Composite of above components	All of the above; baseline measurement before optimisation is essential

The SCI specification is maintained by the Green Software Foundation and is available at https://sci-spec.org/. POSOCO = Power System Operation Corporation. RAPL = Running Average Power Limit (Intel hardware energy measurement). IPMI = Intelligent Platform Management Interface.

The AI Energy Paradox

Training, Inference, and the Carbon Cost of Intelligence

Chapter 7

The AI Energy Paradox — Training, Inference, and the Carbon Cost of Intelligence

There is a paradox at the heart of artificial intelligence's relationship with environmental sustainability, and it demands to be stated plainly before anything else in the present chapter is said. AI systems are, simultaneously, the fastest-growing source of new energy demand in the data centre sector and the most powerful tool available for optimising the energy systems that data centres — and the broader economy — depend upon. Google has used AI to reduce the energy consumed by its data centre cooling systems by 40 per cent. DeepMind's AlphaFold has reduced the computational cost of protein structure prediction by a factor sufficient to compress what might have been decades of experimental work into months. AI systems are being used to optimise the dispatch of renewable energy on constrained electricity grids, to improve the yield forecasting of solar and wind farms, and to detect energy waste in industrial processes that human operators would never notice.

And yet: training the GPT-3 language model is estimated to have consumed approximately 1,300 megawatt-hours of electricity, generating roughly 552 tonnes of carbon dioxide equivalent — more than the lifetime emissions of five average American automobiles. The models that succeeded GPT-3 required substantially more. The inference energy consumed by the ChatGPT service, at its reported peak of one hundred million daily users, has been estimated at ten times the inference energy of an equivalent volume of Google Search queries. India's own growing AI research community — centred at the IITs, IISc, and at private

sector organisations including Infosys, TCS, Wipro, and a rapidly expanding startup ecosystem — is navigating this paradox with the urgency it deserves, yet often without the quantitative tools needed to navigate it wisely.

The present chapter addresses the energy dimensions of AI directly, candidly, and without either the apologetics of the AI industry or the reductive scepticism of its harshest critics.

7.1 The Carbon Cost of Training: What the Evidence Shows

The quantification of the carbon cost of training large artificial intelligence models is a relatively recent area of research, catalysed in large measure by the publication of Strubell, Ganesh, and McCallum's influential 2019 paper Energy and Policy Considerations for Deep Learning in NLP. Their analysis of the energy consumption of training several natural language processing models — including a large transformer model with neural architecture search — produced figures that surprised many in the AI research community and prompted the broader field to begin taking the energy question seriously.

The methodology for estimating training energy and carbon is straightforward in principle, though demanding in practice: one must know the hardware used, the duration of training, the power consumption of that hardware at the relevant utilisation level, and the carbon intensity of the electricity grid supplying the training facility. In practice, all of these quantities are subject to uncertainty: hardware power consumption varies with computational load and thermal state; training runs are rarely completed in a single continuous session; the carbon intensity of cloud computing regions is not always publicly disclosed; and the full scope of overhead energy — cooling, power conversion,

network connectivity — is frequently omitted from published estimates.

Notwithstanding these methodological challenges, the trajectory of AI training energy consumption over the period from 2018 to 2024 is sufficiently well-documented to support a confident observation: it has grown dramatically, at a rate that has substantially outpaced the efficiency improvements achieved through better hardware and algorithms. Patterson and colleagues (2022) estimated the training carbon for a range of milestone models, finding that GPT-3's training emitted approximately 552 tCO_2e, Megatron-LM (530B) approximately 270 tCO_2e, and the Switch Transformer approximately 59 tCO_2e — the lower figure for the latter two reflecting both algorithmic improvements and the use of more carbon-efficient computing infrastructure. The models that have been trained subsequent to the period covered by Patterson et al. — including the GPT-4 family, Claude, Gemini, and the Llama series — have been subject to less transparent disclosure, but available evidence suggests training energy requirements have continued to grow.

It is necessary, at this point, to introduce a distinction that is often elided in public discussion of AI's carbon footprint: the distinction between training energy and inference energy. Training — the process of adjusting a model's parameters to minimise a loss function on a large dataset — is a computationally intensive but one-time cost. Once a model is trained, it need not be retrained to serve users; the trained parameters are fixed and used repeatedly for inference. Inference — the application of a trained model to produce an output for a given input — is computationally much lighter per query than training, but it is executed billions of times, and its aggregate energy cost at the scale of a widely deployed service is potentially far larger than the training cost. For the

ChatGPT service at peak usage, the inference energy was estimated to exceed the training energy within weeks of deployment. For services like Google Search, which have deployed AI components across billions of daily queries, the inference energy now dominates the AI energy budget entirely.

Table 7.1 Carbon Cost of Training Milestone AI Models (2018–2024)

Model	Year	Parameters	Training Energy (MWh, est.)	Training CO_2e (t, est.)	Notes on Methodology
Transformer (Vaswani et al.)	2017	65M	~0.01	~0.005	Small by current standards; estimate extrapolated from hardware spec
BERT-Large	2018	340M	~1.5	~0.65	Strubell et al. (2019) estimate; TPUv3 hardware
GPT-2 (1.5B)	2019	1.5B	~10	~4	OpenAI disclosure; V100 GPU cluster
T5-11B	2019	11B	~87	~47	Strubell et al. extended analysis; TPU cluster, Google infrastructure
GPT-3 (175B)	2020	175B	~1,300	~552	Patterson et al. (2022); V100 cluster, Microsoft Azure
Megatron-LM (530B)	2021	530B	~617	~270	Patterson et al. (2022); A100 cluster, NVIDIA infrastructure
PaLM (540B)	2022	540B	~2,562	~507	Chowdhery et al. (2022); TPUv4

					pods, Google infrastructure
Llama 2 (70B)	2023	70B	~539	~292	Meta AI disclosure; A100 GPUs, AWS infrastructure
Gemini Ultra (est.)	2023	~1T+ (est.)	~3,000–5,000 (est.)	~900–1,500 (est.)	Google disclosure partial; significant uncertainty in estimate

CO_2e figures are estimates with substantial uncertainty; actual values depend on hardware efficiency, data centre PUE, grid carbon intensity, and training methodology specifics. MWh = megawatt-hours. Sources: Strubell et al. (2019); Patterson et al. (2022); Chowdhery et al. (2022); Meta AI Technical Report (2023). Declining CO_2e per parameter for later models reflects improvements in hardware efficiency (A100 vs V100), use of renewable energy, and algorithmic advances.

7.2 Inference at Scale: The Hidden Majority of AI Energy

The drama of large model training — the narrative of vast compute clusters running for months to produce a single model — tends to dominate public and academic discussion of AI's energy footprint. This emphasis is understandable, given the large and striking figures involved, but it is potentially misleading about where the larger share of AI's total energy consumption arises. For any AI service that achieves significant adoption, the energy consumed in inference — applying the trained model to respond to actual user queries — will, within a relatively short period of deployment, exceed the energy consumed in training the model.

The energy consumed per inference query depends on several factors: the size of the model (larger models require more computation per token generated), the length of the input and output (longer contexts and longer generated sequences require more computation), the hardware on which inference is performed,

and the efficiency of the serving infrastructure. For a large language model of the scale of GPT-4, estimates of inference energy per query range from approximately 0.001 to 0.01 kilowatt-hours — a range that reflects substantial uncertainty about model size, hardware configuration, and query characteristics. At one hundred million daily queries, the lower end of this range corresponds to approximately 100 megawatt-hours of daily inference energy — comparable to the energy consumption of a small data centre.

For India, the growth of AI-powered services in high-volume applications creates an inference energy challenge of particular significance. The government's deployment of AI in public service delivery — through the AI-powered grievance resolution system of CPGRAMS, the AI features being integrated into the DigiLocker and UMANG platforms, and the AI-assisted services being developed under the National AI Strategy — involves the execution of inference queries at scales that, taken together, could represent a meaningful fraction of the energy budget of India's public sector IT infrastructure. The design of these systems with inference efficiency as an explicit requirement — rather than as an afterthought — is a matter of both fiscal prudence and environmental responsibility.

7.3 Efficient AI: Quantisation, Pruning, Distillation, and Sparse Models

The recognition that AI model training and inference carry significant energy costs has spurred considerable research into techniques for reducing those costs without proportionate sacrifice of model capability. Four families of technique — quantisation, pruning, knowledge distillation, and sparse models — have

emerged as the most practically significant, and each deserves brief treatment here.

Quantisation reduces the numerical precision used to represent model parameters and activations during inference — or, in some techniques, during training as well. A standard neural network parameter is typically stored as a 32-bit or 16-bit floating-point number. Quantising to 8-bit integers reduces storage and memory bandwidth requirements by a factor of two to four, and on hardware with dedicated integer arithmetic units — including the neural processing units in many mobile SoCs — reduces computation energy by a comparable factor. More aggressive quantisation to 4-bit or even 1-bit representations has been demonstrated with acceptable quality degradation for many tasks, enabling models that would otherwise require data centre hardware to be deployed on smartphones and edge devices. The practical consequence for energy is that quantised models can be served on lower-power hardware, or can perform more inferences per unit of energy on the same hardware.

Pruning removes parameters from a trained model that contribute little to its output quality, reducing the model's size and the computation required for inference. Structured pruning removes entire neurons, attention heads, or layers — producing a smaller model that can be executed more efficiently with standard hardware. Unstructured pruning removes individual weights, producing a sparse model that requires specialised hardware or software to exploit efficiently. The energy savings from pruning are real but depend on the degree of sparsity achieved and the hardware's ability to exploit it.

Knowledge distillation trains a smaller 'student' model to reproduce the outputs of a larger 'teacher' model, producing a compact model with much of the capability of the original at a

fraction of the inference cost. The DistilBERT model, which retains approximately 97 per cent of BERT's performance at 40 per cent of its size and 60 per cent of its inference speed, is a widely cited example of successful distillation. For deployment contexts where inference latency and energy are constraints — mobile devices, edge computing nodes, real-time API endpoints — distilled models offer a practically important trade-off between capability and efficiency.

Sparse Mixture of Experts (MoE) architectures represent perhaps the most structurally interesting approach to efficient large models. Rather than activating all parameters for every input, MoE models maintain a large pool of specialised 'expert' sub-networks and activate only a small fraction of them for any given input, routing each input to the experts most relevant to it through a learned gating mechanism. The Switch Transformer and the Mixtral family of models employ this architecture, achieving the capability of much larger dense models at a fraction of the inference computation — because only a fraction of parameters are active for any given token. The energy efficiency advantage of MoE architectures is substantial for inference: a model with 100 billion total parameters but 10 billion active parameters per token consumes inference energy comparable to a 10-billion-parameter dense model, not a 100-billion-parameter one.

Table 7.2 AI Efficiency Techniques — Energy Reduction and Quality Trade-offs

Technique	Energy Reduction (inference)	Quality Impact	Hardware Requirement	Best Application
FP16 / BF16 mixed precision	~30–50% vs FP32	Negligible (<0.1% quality loss)	Modern GPU/NPU with FP16 support	Standard for all production AI inference;

				essentially no trade-off
INT8 post-training quantisation	~50–70% vs FP32	Minor (0.5–2% quality loss on benchmarks)	Hardware INT8 arithmetic units (TensorRT, ONNX Runtime)	Mobile inference, edge deployment, latency-constrained APIs
INT4 quantisation (GPTQ, AWQ)	~70–80% vs FP32	Moderate (2–5% on some tasks)	Specialised kernels (ExLlama, llama.cpp)	Consumer hardware deployment; acceptable for many chat applications
Structured pruning (50% sparsity)	~40–60%	Moderate — depends on task and fine-tuning after pruning	Standard hardware; no special requirement	When retraining is feasible; structured sparsity suits standard hardware
Knowledge distillation	~40–90% (model-specific)	Task-dependent; typically 3–10% vs teacher	Standard hardware for distilled model	High-volume inference; mobile/edge; real-time applications
Sparse MoE (top-2 of 8 experts)	~60–80% vs equivalent dense model	Competitive with dense models of equivalent active parameters	MoE-aware routing hardware preferred	Large-scale serving where total parameters >> active parameters
Speculative decoding	~30–50% latency reduction	No quality loss (theoretically exact)	Two models: small draft + large verifier	Autoregressive generation; streaming text output

Energy reduction figures are approximate and depend heavily on the specific model, task, hardware, and implementation. Quality impact is measured against the uncompressed baseline model. Sources: Frantar et al. (2022) for GPTQ; Lin et al. (2023) for AWQ; Fedus et al. (2021) for Switch Transformer; Leviathan et al. (2023) for speculative decoding.

7.4 AI for Green: The Beneficial Side of the Equation

The preceding sections of the present chapter have addressed AI's energy consumption as a challenge to be managed. It would be a significant omission, however, to conclude the present discussion without examining the other side of the paradox: the applications of AI to the management and reduction of energy consumption in computing infrastructure, in the electricity grid, and in the broader economy.

Google DeepMind's application of deep reinforcement learning to the management of cooling systems in Google's data centres, reported in 2016 and subsequently extended, is perhaps the most cited example of AI-for-green in computing infrastructure. The system, trained to control the set-points of cooling equipment based on sensor readings and workload predictions, achieved a 40 per cent reduction in cooling energy and a 15 per cent reduction in total data centre PUE compared to the human-operated baseline. The significance of this result lies not only in its magnitude but in its demonstration that the complex, non-linear, time-varying dynamics of a data centre thermal system — a system that is genuinely difficult for human operators to optimise — are amenable to the pattern recognition and optimisation capabilities of deep learning.

In the electricity sector, AI applications relevant to the integration of renewable energy include: improved forecasting of solar and wind generation (which enables grid operators to reduce the reserve capacity they must hold against forecast errors); reinforcement learning-based optimisation of battery storage dispatch (which maximises the economic and environmental value of stored energy); and anomaly detection in transmission and distribution infrastructure (which reduces energy losses from undetected faults). The Central Electricity Authority of India has

initiated programmes to apply machine learning to grid demand forecasting, and the National Smart Grid Mission includes provisions for AI-based grid management tools. These applications, if successfully implemented at the scale of India's national grid, could have energy and carbon implications that dwarf the aggregate training and inference energy of the AI systems used to implement them.

In the scientific domain, the energy implications of AI-accelerated research are potentially the most profound of all. AlphaFold 2's solution to the protein structure prediction problem — a problem that had occupied structural biologists for fifty years — reduced the computational cost of predicting a protein's three-dimensional structure from what might require weeks of molecular dynamics simulation to seconds of neural network inference. The consequence for drug discovery, materials science, and synthetic biology is that experiments that would previously have required expensive and energy-intensive laboratory work can be screened computationally, with only the most promising candidates proceeding to physical synthesis. If AI-accelerated discovery leads to better solar cells, more efficient batteries, or improved catalysts for chemical processes, the energy implications of those discoveries will far exceed the energy consumed in producing them.

Green Cloud Computing

Resource Efficiency, Elasticity, and Sustainability SLAs

Chapter 8

Green Cloud Computing — Resource Efficiency, Elasticity, and Sustainability SLAs

When the Income Tax Department of India migrated its assessment and processing systems to the cloud in phases between 2019 and 2022, the migration was driven primarily by considerations of scalability, resilience, and cost. The department faces a highly seasonal workload: the days immediately preceding the July 31 assessment filing deadline see traffic volumes many times the annual average, followed by months of substantially lower activity. On-premises infrastructure sized for peak demand would sit largely idle for most of the year, consuming electricity for cooling and power conditioning whether or not the servers were performing useful work. Cloud infrastructure, in principle, eliminates this waste: capacity is provisioned for peak demand and released — and its energy consumption reduced — when demand falls.

In practice, the energy efficiency advantage of cloud migration depends critically on how the migration is managed. A simple ‘lift and shift’ — moving virtual machines configured for on-premises operation to cloud instances without re-architecting the applications — often fails to realise the energy savings that cloud infrastructure, in principle, makes possible. The virtual machines continue to be over-provisioned, the applications continue to hold resources they are not using, and the main difference from the energy perspective is that the waste now occurs in someone else’s data centre rather than the department’s own computer room. Realising the energy potential of cloud computing requires not just moving to the cloud but rethinking how

applications are designed, deployed, and managed within it. The present chapter examines what that rethinking entails.

8.1 The Efficiency Promise of Cloud: Consolidation and the Utilisation Gap

The theoretical case for the energy efficiency of cloud computing relative to on-premises data centres rests on two arguments. The first is consolidation: by aggregating the workloads of many organisations onto shared physical infrastructure, cloud providers can achieve server utilisation rates substantially higher than individual organisations operating their own facilities. The second is specialisation: cloud providers, whose core business is the efficient operation of computing infrastructure, have the scale, expertise, and incentive to invest in the most energy-efficient hardware, cooling systems, and operational practices available — investments that most individual organisations cannot justify.

The empirical evidence for these arguments is reasonably compelling for large-scale migration from inefficient on-premises infrastructure to well-managed cloud services. A study by Lawrence Berkeley National Laboratory estimated that migrating US enterprise workloads to the cloud could reduce the energy consumed by those workloads by approximately 87 per cent — primarily through the elimination of idle server capacity and the improvement in utilisation rates from a typical enterprise average of 12 to 15 per cent to the hyperscale average of 65 to 80 per cent. The efficiency of hyperscale data centres — with PUEs averaging 1.1 to 1.2 compared to the enterprise average of 1.5 to 2.0 — contributes additional savings.

However, the empirical evidence also reveals that this efficiency promise is not automatically realised. The phenomenon

of cloud sprawl — the uncontrolled proliferation of cloud resources that are provisioned for convenience and never fully utilised — has been extensively documented in enterprise cloud deployments. A 2023 study by Flexera found that 32 per cent of enterprise cloud spending was estimated to be wasted — on over-provisioned instances, unused storage, and zombie resources that were provisioned and forgotten. From an energy perspective, a cloud instance that is provisioned but idle still consumes electricity — approximately 20 to 40 per cent of its peak consumption in most instance types — and the waste is dispersed across the cloud provider's infrastructure rather than concentrated in the organisation's own computer room, making it invisible to the organisation's energy accounting.

For Indian enterprises and government organisations adopting cloud services — whether through MeitY's Government Cloud (GI Cloud/MeghRaj), the National Informatics Centre's cloud infrastructure, or commercial cloud services from AWS, Azure, Google Cloud, and Indian providers such as NIC, ESDS, and Tata Communications — the utilisation gap represents the most readily addressable source of energy waste. The discipline of FinOps — cloud financial operations — which applies cost optimisation practices to cloud resource management, is simultaneously a cost optimisation and an energy optimisation, because the charges that cloud providers levy for compute resources are, in the main, proportional to the energy those resources consume.

8.2 Virtualisation, Containers, and Serverless: An Energy Hierarchy

The fundamental technology that enables cloud computing's consolidation efficiency is virtualisation — the sharing of physical hardware among multiple isolated workloads, each believing it has

exclusive access to its own server. Virtualisation introduces an overhead — the hypervisor that manages the virtual machines consumes resources — but this overhead is typically small compared to the consolidation benefit it enables: a physical server that previously ran a single workload at 15 per cent utilisation can, after virtualisation, run ten workloads, each at 15 per cent utilisation of its virtual allocation, achieving an aggregate physical utilisation of 60 to 80 per cent.

Container technology — exemplified by Docker and managed through orchestration platforms such as Kubernetes — provides a lighter-weight isolation mechanism than full virtualisation, sharing the operating system kernel among multiple isolated application environments rather than virtualising the complete hardware stack. The energy overhead of containers is smaller than that of virtual machines: a container-based deployment typically requires 10 to 40 per cent less memory and has lower per-request latency than an equivalent VM-based deployment, both of which translate into energy savings at scale. For microservices architectures — in which a large application is decomposed into many small, independently deployable services — container orchestration enables the fine-grained scaling of individual services in response to demand, further improving resource utilisation.

Serverless computing — the Function as a Service (FaaS) model offered by AWS Lambda, Azure Functions, Google Cloud Functions, and similar services — takes the resource efficiency principle to its logical extreme. In the serverless model, the cloud provider allocates resources to execute a function only when that function is invoked, releasing those resources immediately upon completion. For workloads that are invoked intermittently — event-driven processing, API backends with variable traffic, scheduled batch jobs — serverless eliminates the idle resource

consumption that persists even in well-managed container deployments. The energy cost of serverless execution is, in principle, proportional to the execution itself, rather than to the continuous allocation of resources to handle potential demand.

The qualification ‘in principle’ in the preceding sentence requires elaboration. Serverless functions incur a ‘cold start’ penalty when they are invoked after a period of inactivity: the cloud provider must allocate resources, load the runtime environment, and initialise the function before it can serve the request, a process that takes tens to hundreds of milliseconds and consumes energy not counted in the function’s own execution time. Cold starts represent a real, if often overlooked, component of serverless energy consumption. Additionally, the very ease with which serverless functions can be created and invoked creates a risk of serverless sprawl analogous to cloud sprawl: many small functions consuming collectively significant energy in aggregate, with no single function large enough to attract the attention of resource optimisation efforts.

Table 8.1 Cloud Deployment Models — Energy Efficiency Comparison

Deployment Model	Typical Server Utilisation	Overhead (vs bare metal)	Idle Energy (% of peak)	Energy Proportionality	Best For
Bare metal (on-premises, legacy)	~10–15%	None	~60–70%	Poor — nearly constant power	High-security, latency-critical; legacy applications with specific hardware needs

Virtualised VMs (on-premises)	~30–50%	Hypervisor ~5–10%	~40–60%	Poor–Moderate — VMs sized at allocation	Enterprise workloads; existing on-premises infrastructure modernisation
Cloud IaaS (virtual machines)	~40–70% (provider fleet)	Hypervisor ~5%	~20–40% of instance peak	Moderate — scales by instance type	General enterprise; lift-and-shift migration; predictable workloads
Cloud containers (Kubernetes)	~50–80% (pod level)	~3–5% runtime overhead	~10–20% of container peak	Good — per-pod scaling	Microservices; cloud-native applications; dev/test environments
Serverless / FaaS	~90%+ (execution periods)	Cold start ~50–200 ms per invocation	~0% between invocations	Very Good — charges for execution only	Event-driven; intermittent; API handlers; unpredictable traffic patterns
Container + KEDA autoscaling (scale-to-zero)	~85–95% during active periods	KEDA controller overhead minimal	~0% when scaled to zero	Excellent — container efficiency with serverless idle behaviour	Batch jobs; queue-driven workloads; development environments

Utilisation figures represent the achievable range under well-managed deployments; actual figures vary significantly with workload characteristics, organisation maturity, and provider configuration. IaaS = Infrastructure as a Service. FaaS = Function as a Service. KEDA = Kubernetes Event-Driven Autoscaling. Sources:

Lawrence Berkeley National Laboratory (2023); Flexera State of the Cloud Report (2023); Google SRE Book; AWS Well-Architected Framework.

8.3 Carbon-Aware Workload Placement: The Geographic Dimension

The preceding sections of the present chapter have addressed the energy efficiency of cloud infrastructure primarily in terms of hardware utilisation and deployment model. An equally important dimension — and one that has received growing attention in the research literature and in the practices of leading cloud operators — is the geographic placement of workloads, which determines the carbon intensity of the electricity used to power them.

The carbon intensity of electricity varies substantially across the world's electricity grids, from near-zero for grids powered predominantly by hydroelectricity, nuclear energy, or renewable sources, to over 800 grammes of CO_2 equivalent per kilowatt-hour for grids dominated by coal-fired generation. A computation performed in Iceland — where geothermal and hydroelectric power provide nearly 100 per cent of electricity at effectively zero carbon intensity — generates a small fraction of the carbon of the same computation performed in a region where coal is the dominant source. For cloud computing workloads that are flexible in their geographic placement — that do not require data to remain in a specific jurisdiction, that can tolerate the small latency penalty of geographic distance — this variation represents an opportunity for substantial carbon reduction without any sacrifice of computational result.

The Google Cloud Carbon Footprint tool and Microsoft Azure's Emissions Impact Dashboard represent the current state of practice in making geographic carbon intensity data available to cloud customers, enabling them to evaluate the carbon implications

of their resource placement choices. Both tools provide region-level carbon intensity estimates based on the electricity grid composition of each cloud region, updated periodically if not in real time. The practical challenge of acting on this information is that workload placement decisions in cloud architectures are often made implicitly — through default region selection in deployment scripts — rather than explicitly, and the organisational processes for incorporating carbon considerations into cloud architecture decisions are not yet well-established in most organisations.

For Indian cloud customers, the carbon-aware placement question has a specific and important dimension: the question of data residency. India's data protection regulations, and the operational requirements of many applications, mandate that certain categories of data remain within India's geographic boundaries. This constraint limits the degree to which Indian cloud workloads can be geographically shifted to exploit lower-carbon grids elsewhere. Within India, however, meaningful variation in grid carbon intensity exists between regions — as was noted in Chapter 6 — and the selection of cloud regions in Tamil Nadu, Karnataka, or Kerala over regions in Uttar Pradesh or Bihar can result in materially lower carbon emissions for comparable workloads. As India's renewable energy capacity continues to grow and geographic variation in grid carbon intensity decreases, this consideration will become less acute, but it remains relevant for the current decade.

8.4 Sustainability SLAs: The Emerging Standard of Cloud Accountability

The concept of a Service Level Agreement (SLA) — the contractual specification of the performance, availability, and reliability that a cloud provider commits to deliver, with financial

penalties for non-compliance — is well-established in the cloud computing industry. What is considerably less well-established, but is beginning to emerge in the practices of leading providers and in the demands of sophisticated customers, is the sustainability SLA: the extension of the SLA concept to cover the environmental performance of cloud services, including energy efficiency, carbon intensity, water consumption, and hardware lifecycle management.

The barriers to the development of sustainability SLAs are both technical and commercial. On the technical side, the measurement and attribution of energy consumption and carbon emissions to individual customer workloads in a shared cloud environment is a genuinely difficult problem: the energy consumed by a virtual machine is not easily separable from the energy consumed by the other virtual machines sharing the same physical host, the cooling system serving the rack, or the power distribution infrastructure serving the facility. Cloud providers have developed attribution methodologies that allocate a proportional share of facility energy to each workload, but these methodologies differ among providers in ways that make comparisons difficult. On the commercial side, providers have historically been reluctant to accept contractual commitments on environmental metrics that are subject to factors — grid carbon intensity, weather-dependent cooling efficiency — partly beyond their control.

Notwithstanding these barriers, progress is being made. The Science Based Targets initiative's Corporate Net-Zero Standard, which requires companies to set emissions reduction targets consistent with limiting global warming to 1.5 degrees Celsius, has prompted a growing number of major corporations to demand verifiable emissions data from their cloud providers as part of their Scope 3 accounting. The EU's Corporate Sustainability Reporting Directive, which requires large companies and listed entities to

disclose detailed climate-related information, creates a direct demand for supply chain emissions data that cloud providers must ultimately supply. In India, the Business Responsibility and Sustainability Reporting (BRSR) framework, which became mandatory for the top 1,000 listed companies in India from the financial year 2022–2023, includes requirements for disclosure of energy consumption and greenhouse gas emissions that extend to supply chain emissions for the BRSR Core category — creating a domestic driver for cloud provider sustainability accountability that will strengthen over time.

The present chapter closes with the observation that the path to genuinely green cloud computing runs through accountability as much as through technology. The hardware efficiency, cooling innovation, and renewable energy procurement addressed in earlier chapters provide the technical foundation; but without measurement standards, disclosure requirements, and procurement practices that reward environmental performance, the incentives for providers to invest in sustainability will remain weaker than the incentives to compete on price and performance. The regulatory and procurement landscape in India is, as the preceding paragraphs indicate, beginning to create those incentives. The speed with which the cloud industry responds to them will be, in significant measure, a function of how consistently customers — including the Government of India in its own procurement — make sustainability a criterion of consequence in cloud provider selection.

Table 8.2 Cloud Provider Sustainability Commitments and Reporting (as of 2024)

Provider	Net-Zero Commitment	Renewable Energy (2023)	Carbon Reporting Tool	Water Commitment	India-Specific Disclosure
Amazon Web Services (AWS)	Net-zero by 2040 (Amazon Climate Pledge)	~100% matched annually; 90% hourly in some regions	AWS Customer Carbon Footprint Tool (Scope 1, 2, 3 partial)	Water positive by 2030	Limited; India regions not individually disclosed
Microsoft Azure	Carbon negative by 2030; historical removal by 2050	100% matched; 24/7 CFE in progress	Azure Emissions Impact Dashboard; detailed Scope 3	Water positive by 2030	India region carbon intensity reported; water data limited
Google Cloud	Net-zero by 2030; 24/7 CFE by 2030	100% matched since 2017; 64% 24/7 CFE (2023)	Google Cloud Carbon Footprint; hourly carbon data (beta)	Water replenishment target 120% by 2030	India region included in carbon footprint tool; improving disclosure
NIC / MeghRaj (Government of India)	Under development; not yet committed	Mix of state grid (primarily coal-dependent)	Not available publicly	Not committed	Full disclosure not available; BEE assessment ongoing
ESDS / CtrlS / Tata Communications (Indian private)	Vary; some RE commitments; no net-	Varies; some solar PPAs; predomina	Limited public disclosure	Limited	Limited; voluntary sustainability reports by

	zero timelines	ntly grid-powered			some operators

Commitments and reporting tools are subject to rapid change; readers should verify current status from provider sustainability reports. NIC = National Informatics Centre. CFE = Carbon-Free Energy. BRSR = Business Responsibility and Sustainability Reporting. The absence of comprehensive sustainability disclosure from Indian cloud providers represents a significant gap in the accountability framework; the BRSR requirement for supply chain emissions disclosure is expected to drive improvement. Sources: AWS, Azure, and Google sustainability reports (2024); NIC Annual Report (2023); SEBI BRSR guidelines (2022).

Measuring Green

Standards, Metrics, Reporting, and the Greenwashing Problem

Chapter 9

Measuring Green — Standards, Metrics, Reporting, and the Greenwashing Problem

In 2021, a major European cloud provider published a sustainability report stating that its operations were ‘powered by 100 per cent renewable energy.’ The claim was accompanied by a striking graphic of wind turbines and solar panels. It was, in a narrow technical sense, accurate: the company had purchased renewable energy certificates (RECs) in a quantity matching its annual electricity consumption. What the report did not disclose was that the renewable energy in question had been generated in a different country, at a different time of year, and fed into a different electricity grid than the one supplying the company’s data centres. The company’s servers, running around the clock in a grid region whose overnight and winter electricity came predominantly from fossil fuel generation, had consumed that fossil fuel generation without interruption. The renewable energy existed; it powered other users in other places at other times. The claim was not false. It was, however, profoundly misleading.

This episode is far from unique. It illustrates a phenomenon that has become endemic in the sustainability reporting of the technology industry and that poses a genuine obstacle to the progress of green computing: greenwashing — the presentation of environmental claims that are technically defensible but substantively misleading about an organisation’s actual environmental impact. The present chapter examines the measurement frameworks, reporting standards, and regulatory instruments that have been developed to distinguish genuine environmental performance from its appearance, and the work that

remains to be done — in India and globally — before that distinction can be made reliably and at scale.

9.1 The Measurement Problem: Why Green Computing Is Hard to Quantify

The fundamental challenge in measuring the environmental performance of computing is not the absence of metrics — there is, if anything, a surfeit of them — but the absence of a single, widely adopted, and genuinely comprehensive metric that captures all relevant dimensions of impact in a form that is comparable across organisations, technologies, and time periods. The Power Usage Effectiveness (PUE) metric, introduced in Chapter 4, is the most widely adopted of the available metrics, but as was noted there, it captures only one dimension of a multidimensional problem. A facility with an excellent PUE may be located on a coal-heavy grid, may consume extraordinary quantities of water in its cooling systems, and may refresh its hardware on an unnecessarily short cycle that generates significant embodied carbon. None of these dimensions is captured by PUE.

The measurement problem is compounded by the structure of the IT supply chain, which is among the most globally distributed and opaque of any major industry. The components of a server purchased by an Indian bank or government ministry may have been designed in the United States, manufactured on wafers fabricated in Taiwan, assembled in Malaysia, fitted with hard disk drives made in Thailand, and shipped via Singapore. The carbon embedded in this supply chain — the emissions generated at each step of design, fabrication, assembly, and shipping — is distributed across multiple countries, multiple corporate entities, and multiple accounting frameworks. No single company in this chain has

visibility into the full picture, and no existing accounting standard compels them to provide it.

These structural challenges do not make measurement impossible; they make it difficult and require careful methodological choices. The present section surveys the principal measurement frameworks available to practitioners and researchers, noting their scope, their limitations, and their relevance to the Indian context.

9.2 The Metrics Landscape: A Critical Survey

The reader who has followed the preceding chapters will have encountered several of the major green computing metrics in context: PUE, WUE, CUE, ERE, and SCI have each been introduced and discussed in the chapters to which they are most directly relevant. The purpose of the present section is to assemble these metrics into a coherent comparative framework and to add several that have not yet been addressed — particularly those relevant to life cycle assessment and organisational carbon accounting.

Life Cycle Assessment (LCA) is the most comprehensive methodology available for quantifying the environmental impact of a product, process, or service across its entire life cycle — from the extraction of raw materials through manufacturing, distribution, use, and end-of-life treatment. An LCA of a server, conducted in accordance with ISO 14040 and 14044, would account for the energy and emissions associated with mining the metals and rare earth elements in its components, the fabrication of its semiconductor chips, the energy consumed during its operational life, and the energy and material recovery at end of life. LCA is methodologically rigorous and comprehensive by design, but it is also time-consuming and data-intensive, and its results are

sensitive to methodological choices — the system boundary, the allocation of shared impacts, the characterisation of environmental midpoints — that make comparisons between independently conducted LCAs difficult.

The Greenhouse Gas Protocol — developed jointly by the World Resources Institute and the World Business Council for Sustainable Development and first published in 2001, with subsequent revisions — provides the most widely adopted framework for corporate greenhouse gas accounting. Its three-scope structure, introduced in Chapter 1, organises emissions into direct (Scope 1), indirect from electricity (Scope 2), and value chain (Scope 3) categories, and its methodological guidelines provide the basis for the disclosures required by most major voluntary and mandatory reporting frameworks. For an Indian IT company, GHG Protocol accounting would cover the diesel consumed in its own generators (Scope 1), the electricity consumed in its owned or leased facilities (Scope 2), and the emissions from its hardware supply chain, business travel, employee commuting, and the use and disposal of its products (Scope 3).

It is pertinent to note, in the context of the Indian regulatory landscape, that the Securities and Exchange Board of India's Business Responsibility and Sustainability Reporting (BRSR) framework represents the most consequential domestic development in corporate sustainability reporting of recent years. Mandatory for the top 1,000 listed companies from FY 2022–2023 and with a BRSR Core subset requiring assurance from FY 2023–2024, it brings India into alignment with international sustainability reporting norms in a manner that has direct implications for the IT sector's environmental accountability. The National Stock Exchange's inclusion of sustainability disclosure in

its listing requirements, and the alignment of BRSR with the International Sustainability Standards Board's IFRS S1 and S2 frameworks — the emerging global standard for sustainability-related financial disclosures — means that Indian listed companies increasingly face a converging set of domestic and international reporting obligations that make comprehensive environmental measurement a practical necessity rather than an optional exercise.

Table 9.1 Green Computing Metrics — Comprehensive Comparative Reference

Metric	Full Name	Formula / Basis	Scope	Strengths	Limitations	Indian Regulatory Status
PUE	Power Usage Effectiveness	Total facility power ÷ IT equipment power	Facility-level operational energy	Simple; widely adopted; enables benchmarking	Ignores carbon intensity, utilisation, water, embodied carbon	Recommended by BEE; voluntary
WUE	Water Usage Effectiveness	Litres water consumed ÷ IT equipment kWh	Direct facility water consumption	Captures critical water resource dimension	Omits indirect water in grid electricity; not standardised	Not yet required; gap in regulation
CUE	Carbon Usage Effectiveness	$kgCO_2e$ emitted ÷ IT equipment kWh	Operational carbon (Scope 1 + 2)	Captures grid carbon intensity	Omits Scope 3; snapshot metric	Voluntary; BRSR Scope 2 partially covers

ERE	Energy Reuse Effectiveness	(Total − reused energy) ÷ IT equipment kWh	Facility with waste heat reuse	Credits heat recovery; encourages circularity	Requires third-party verification; uncommon	Not reported in India
SCI	Software Carbon Intensity	(E × I + M) per functional unit	Software-level operational + embodied	Software-level accountability; supports optimisation	New standard; tooling immature; comparisons difficult	Emerging; Green Software Foundation
LCA	Life Cycle Assessment	ISO 14040/14044 inventory and impact analysis	Full product lifecycle	Most comprehensive; captures embodied carbon	Time-intensive; data-scarce; comparison-difficult	Not mandatory; referenced in EPR rules
GHG Protocol (S1+S2+S3)	Greenhouse Gas Protocol	Scope 1/2/3 inventory per WRI/WBCSD standard	Organisational value chain emissions	Comprehensive; internationally recognised baseline	Scope 3 quality varies; self-reported; verification optional	BRSR requires Scope 1+2; S3 in Core
TCFD	Task Force on Climate-related Financial Disclosures	Governance, strategy, risk, metrics framework	Climate risk disclosure to investors	Links sustainability to financial materiality	Qualitative; governance-focused; not energy-specific	Voluntary; SEBI encourages alignment

BEE = Bureau of Energy Efficiency. BRSR = Business Responsibility and Sustainability Reporting (SEBI). WRI = World Resources Institute. WBCSD = World

Business Council for Sustainable Development. EPR = Extended Producer Responsibility. IFRS S1/S2 = International Sustainability Standards Board disclosure standards. Sources: The Green Grid (2022); GHG Protocol (2023); Green Software Foundation SCI Specification (2023); SEBI BRSR Guidelines (2022); ISO 14040 (2006).

9.3 Greenwashing in the Technology Sector: Patterns and Detection

The term ‘greenwashing’ — the communication of environmental claims that are misleading, exaggerated, or unsubstantiated — is applied to a broad spectrum of behaviour, from deliberate deception to inadvertent miscommunication resulting from the genuine complexity of environmental accounting. In the technology sector, greenwashing takes several characteristic forms, each enabled by specific gaps or ambiguities in the measurement and reporting frameworks described in the preceding section.

Renewable energy certificate washing — illustrated in the opening vignette of the present chapter — is perhaps the most pervasive form of technology sector greenwashing. The purchase of RECs or guarantees of origin in a quantity matching annual energy consumption allows an organisation to claim that its operations are ‘powered by 100 per cent renewable energy,’ regardless of whether the renewable energy in question was generated in the same grid, at the same time, or even in the same country as the electricity actually consumed. The International Renewable Energy Agency and the Rocky Mountain Institute have both published analyses demonstrating that REC-based claims systematically overstate the decarbonisation achieved by the organisations making them, because RECs can be purchased from renewable projects that would have been built regardless of the purchase, generating no additive climate benefit. The 24/7 carbon-

free energy standard described in Chapter 4 represents the methodologically rigorous alternative to REC-based annual matching, and its adoption by Google and, increasingly, by other hyperscale operators represents genuine progress in the quality of renewable energy claims.

Selective metric reporting is a related pattern in which organisations prominently disclose metrics on which they perform well — PUE for operators of efficient facilities, renewable energy percentage for operators who purchase RECs aggressively — while omitting metrics on which they perform less well. A data centre operator that reports an excellent PUE while declining to disclose server utilisation rates, water consumption, or hardware refresh cycles is providing a partial picture that creates a misleading overall impression. The proliferation of voluntary reporting frameworks, each covering a different subset of environmental impacts, enables this selective disclosure by allowing organisations to choose which framework to report against based on where their performance is strongest.

Scope 3 omission is the most consequential form of greenwashing by magnitude: the reporting of Scope 1 and Scope 2 emissions while omitting Scope 3, which typically represents 70 to 90 per cent of the total. An Indian IT services company that reports impressive reductions in the electricity consumption of its office buildings and data centres, while omitting the emissions embedded in the cloud services it consumes, the hardware it purchases, and the business travel its employees undertake, is disclosing a fraction of its true footprint in a manner that creates a significantly distorted picture of its environmental performance. The BRSR Core requirements, which mandate assurance for Scope 3 emissions for India's largest listed companies from FY 2023–2024, represent a meaningful step toward closing this gap, though the quality and

completeness of Scope 3 reporting will take several years to improve from the current baseline.

Needless to emphasise, the purpose of identifying these patterns is not to impugn the integrity of any particular organisation, but to equip the reader — whether as a researcher, a regulator, a procurer, or an informed citizen — with the analytical tools needed to evaluate environmental claims critically and to demand the quality of evidence that those claims require. The green computing community has both the technical competence and the professional obligation to raise the standard of environmental measurement and accountability in the technology sector. The chapters of the present volume have been written in the conviction that doing so is among the most important contributions that engineers and computer scientists can make to the challenge of sustainable development.

Table 9.2 Global Regulatory Framework for ICT Sustainability Reporting (as of 2024)

Jurisdiction / Framework	Scope	Mandatory / Voluntary	Key ICT Provisions	Status in India
EU Corporate Sustainability Reporting Directive (CSRD)	Large EU companies + non-EU companies with significant EU operations	Mandatory (phased from 2024)	Full ESRS disclosure including Scope 3 supply chain; double materiality assessment	Affects Indian IT exporters to EU; compliance required for EU-listed subsidiaries
EU Carbon Border Adjustment Mechanism (CBAM)	Selected carbon-intensive imports to EU	Mandatory (transition from 2023)	Initially covers steel, cement, aluminium — not ICT; expansion to	Monitor for expansion; preparatory Scope 3 data needed

			electronics possible	
SEC Climate Disclosure Rule (USA)	US-listed companies	Mandatory (phased, partially stayed)	Scope 1 + 2 disclosure; Scope 3 if material; climate risk governance	Applies to Indian companies listed on US exchanges (ADRs)
ISSB IFRS S1 / S2	Global voluntary standard; adopted by several jurisdictions	Voluntary globally; mandatory in some countries	Climate risk and opportunity disclosure; TCFD-aligned	SEBI encouraging alignment; likely to become reference for BRSR evolution
India BRSR (SEBI)	Top 1,000 NSE/BSE listed companies	Mandatory from FY 2022–23; Core assurance from FY 2023–24	Scope 1+2 required; Scope 3 in Core; energy, water, waste KPIs	Domestic baseline; improving quality expected over 3–5 years
India E-Waste Management Rules (2022)	Producers, importers, refurbishers of electronics	Mandatory	Extended Producer Responsibility (EPR) with deposit-refund mechanism	Enforced by CPCB; compliance improving gradually
ISO 14064 / 14001	Any organisation voluntarily	Voluntary; required by some procurement frameworks	GHG inventory methodology; environmental management system	Adopted by leading Indian IT companies for certification; not mandatory

ESRS = European Sustainability Reporting Standards. CBAM = Carbon Border Adjustment Mechanism. ISSB = International Sustainability Standards Board. ADR = American Depositary Receipt. CPCB = Central Pollution Control Board. The regulatory landscape is evolving rapidly; readers should verify current status of each framework before reliance. Sources: European Commission; SEC; ISSB; SEBI; Ministry of Environment, Forest and Climate Change.

The Road to Net-Zero Computing

Emerging Technologies, Policy Levers, and Open Frontiers

Chapter 10

The Road to Net-Zero Computing — Emerging Technologies, Policy Levers, and Open Frontiers

On the fourteenth of September 2023, IBM Research released a technical paper describing a chip called NorthPole. The paper, published in Science, reported that NorthPole — a neuromorphic processor designed for AI inference — achieved an energy efficiency for image recognition tasks that was, on the paper's own figures, 25 times better than a comparable GPU. It accomplished this not by using newer or more exotic materials, not by operating at lower temperatures that would require expensive refrigeration, and not by departing from the silicon manufacturing processes that the global semiconductor industry has been refining for seventy years. It accomplished this by fundamentally rethinking the architecture of computation: by eliminating the separation between processing and memory that, as Chapter 2 established, is responsible for a dominant share of current processor energy consumption, and by drawing inspiration from the architecture of the biological brain — which has, over hundreds of millions of years of evolution, solved the problem of efficient computation with a sophistication that our best silicon systems have not yet approached.

That single result — 25 times better efficiency on a commercially relevant workload, achieved on manufacturable silicon — contains more cause for genuine optimism about the long-term energy trajectory of computing than any quantity of renewable energy certificates. It is a demonstration that the laws of physics do not forbid dramatic improvement; that the gap between current practice and physical limit is real and crossable; and that

the research investment required to cross it is not beyond the reach of the human institutions that have the responsibility to make it.

The present chapter examines the technologies, policy instruments, and research directions that define the road from where computing stands today to where it needs to be. It is written not in a spirit of complacency — the challenges are real and the timeline is urgent — but in a spirit of informed, evidence-grounded conviction that the challenge is surmountable, and that the students and researchers who read the present volume will play a consequential role in surmounting it.

10.1 Neuromorphic Computing: Learning from the Brain

The human brain is, by any reasonable measure, the most energy-efficient computing system known to exist. It performs an extraordinary range of cognitive tasks — perception, language, reasoning, memory, motor control — consuming approximately 20 watts of power, a figure that a single modern GPU can exceed during a single matrix multiplication. The brain achieves this efficiency through an architecture radically different from the von Neumann architecture that underlies virtually all conventional computers: it tightly integrates computation and memory in the same physical structures (neurons and synapses), processes information in parallel and in an event-driven manner rather than sequentially and synchronously, and represents information in sparse patterns of spikes rather than in continuous floating-point values.

Neuromorphic computing is the engineering discipline that attempts to realise, in silicon and related materials, some of the energy efficiency principles of the biological neural architecture. The term was coined by Carver Mead of Caltech in the 1980s, but the field has gained significant momentum in the past decade as

the limitations of conventional architectures have become pressing and as advances in very large-scale integration have made the fabrication of complex neuromorphic circuits more accessible.

Intel's Loihi processor, introduced in 2017 and significantly enhanced in Loihi 2 in 2021, represents the most widely studied neuromorphic research platform currently available. Loihi implements a spiking neural network architecture in which artificial neurons communicate through discrete events (spikes) rather than continuous values, and in which synaptic weights are stored locally alongside the neurons they connect — eliminating the need for repeated large-scale weight fetches from external memory. For tasks well-suited to its architecture — including sparse pattern recognition, optimisation problems, and adaptive control — Loihi achieves energy efficiencies substantially better than equivalent GPU-based implementations.

IBM's NorthPole, described in the opening vignette, takes a different but related approach: rather than implementing biologically realistic spiking dynamics, it focuses on eliminating off-chip memory access by placing 256 megabytes of on-chip SRAM adjacent to the compute cores, so that inference for models that fit within this on-chip memory requires no DRAM access at all. The result is a processor whose energy consumption is dominated by arithmetic rather than memory access — the inverse of the situation described in Chapter 2 for conventional inference hardware — and whose energy efficiency for the targeted inference workloads is correspondingly excellent.

It is pertinent to note, in the context of Indian research and development, that neuromorphic computing represents an area of genuine strategic opportunity. The design of neuromorphic processors does not require access to the most advanced fabrication nodes — NorthPole was fabricated at Samsung's 12-nanometre

process, a technology accessible to Indian-partnered foundry agreements, rather than at the 3-nanometre cutting edge. The research community at IIT Bombay's Department of Electrical Engineering, IIT Delhi's Bharti School of Telecommunication Technology and Management, and the Centre for Nano Science and Engineering at IISc has developed expertise in low-power circuit design and neuromorphic architectures that provides a foundation for more directed investment in this domain.

10.2 Optical and Photonic Computing: Light as the Computational Medium

One of the fundamental sources of energy dissipation in electronic circuits is resistance: when electrons travel through a conductor, they collide with the atoms of the conductor and transfer kinetic energy to them as heat. Photons — the quanta of light — do not carry electric charge and do not experience this resistive dissipation. Optical signal transmission, which uses photons rather than electrons as the carrier of information, is capable of moving information at the speed of light over virtually lossless optical fibres. If computation itself could be performed optically — if the switching and routing operations that constitute the logic of a computer could be implemented in photonic rather than electronic devices — the resistive dissipation of conventional electronics could be eliminated, at least in principle.

Photonic computing is not a new idea; optical computing research has been pursued intermittently since the 1980s. What has changed in the past decade is the maturation of silicon photonics — the integration of photonic components (waveguides, modulators, detectors, ring resonators) onto silicon substrates using processes compatible with conventional semiconductor manufacturing. Silicon photonics enables the fabrication of

photonic circuits at the scale and cost of conventional integrated circuits, making it feasible for the first time to integrate photonic computing elements into practical systems.

The most immediately practical application of photonic computing is in the acceleration of matrix multiplication — the dominant operation in neural network inference — using optical interference. An optical matrix-vector multiplier uses a network of Mach-Zehnder interferometers, each implementing a programmable phase shift that corresponds to a matrix element, to multiply a vector of optical signals by a matrix of coefficients in a single pass through the photonic circuit, consuming energy only in the modulators and detectors rather than in arithmetic circuits. Lightmatter, a US startup, and several academic research groups have demonstrated photonic matrix accelerators capable of performing matrix-vector multiplication at energies per operation substantially below those of GPU arithmetic units, at optical communication speeds.

Photonic computing faces important practical challenges that have thus far prevented its widespread deployment. Programming a photonic circuit requires setting the phase shifts of many interferometers with high precision, and small thermal and fabrication variations can cause the intended phase to drift from the actual phase, introducing computation errors. The precision engineering required to address these challenges adds cost and complexity. Additionally, photonic circuits currently require electro-optical conversion at their interfaces with conventional electronics — the optical signals must be converted from and to electronic signals at the boundaries of the photonic computing region — and these conversions consume energy and add latency. The research and engineering required to make photonic computing practical at scale is substantial, and realistic timelines

for commercial deployment of general-purpose photonic processors are measured in decades rather than years. It is, however, a direction that the present author regards with genuine scientific excitement, and one that Indian research institutions — particularly those with strengths in photonics, integrated optics, and signal processing, such as IIT Madras and the Raman Research Institute — are well-positioned to contribute to.

10.3 Quantum Computing: Potential and the Energy Reality

Quantum computing receives considerable attention in discussions of future computing technology, and some portion of that attention is justified by the genuine and remarkable capabilities that quantum algorithms theoretically offer for specific problem classes. Shor's algorithm for integer factorisation, Grover's algorithm for unstructured database search, and variational quantum algorithms for quantum chemistry simulation offer computational advantages over the best known classical algorithms that range from quadratic to exponential. For problems in these specific classes — of which the simulation of quantum systems for drug discovery and materials science is among the most practically significant — quantum computers of sufficient scale and quality could offer transformative capability.

The energy dimension of quantum computing is, however, considerably more complicated than the public discourse typically acknowledges, and it is an obligation of intellectual honesty to address it carefully. Current quantum computers — using superconducting qubits of the kind employed by IBM, Google, and the Indian government-supported quantum computing initiative at C-DAC — require the qubits to be maintained at temperatures approaching 15 millikelvin, approximately a hundred times colder

than the temperature of deep space. Achieving and maintaining these temperatures requires dilution refrigerators that consume kilowatts to tens of kilowatts of power, cooling a processor that itself contains only a few hundred to a few thousand qubits and performs computations equivalent to what a classical computer can accomplish with far less energy.

At the present state of the art, quantum computers are not energy-efficient by any meaningful measure; they are extraordinary scientific instruments that demonstrate principles of profound importance but that do not yet constitute an energy-efficient computing technology. The path from current quantum computers — noisy intermediate-scale quantum (NISQ) devices with hundreds of qubits and high error rates — to fault-tolerant quantum computers of the scale needed to execute Shor's algorithm on practically relevant problem sizes requires advances in qubit quality, error correction, and, critically, the overhead of qubit control electronics and refrigeration that are not currently addressed in most public discussions. The National Mission on Quantum Technologies and Applications, launched by the Government of India with an allocation of Rs 6,000 crore, represents a substantial commitment to developing quantum computing capability; the energy efficiency of quantum computing architectures should be an explicit consideration in the research agenda it supports.

Table 10.1 Emerging Compute Paradigms — Energy Promise, Maturity, and Indian Relevance

Paradigm	Energy Promise	Technical Maturity (2024)	Time to Commercial Scale	Indian Research Strength	Strategic Priority for India
Neuromorphic computing (spiking NN)	10–100× better than GPU for inference	Research / limited commercial (Intel Loihi 2, IBM NorthPole)	5–10 years for niche applications	IIT Bombay, IISc CENSE — circuit design expertise	High — manufacturable at accessible nodes; IoT and defence applications
Neuromorphic computing (in-memory, analogue)	100–1,000× for specific tasks	Early research phase; key challenges in precision and reliability	10–15 years	Emerging groups at IIT Delhi, IIT Hyderabad	Medium — high risk; high reward if precision challenges solved
Silicon photonic matrix accelerators	5–50× vs GPU for matrix-heavy inference	Early commercial demonstrations (Lightmatter, Luminous Computing)	7–12 years at scale	IIT Madras photonics; Raman Research Institute	Medium — IIT Madras photonics group well-positioned
Free-space optical interconnects	Reduction in chip-to-chip communication energy	Advanced research; some deployment in data centre interconnect	3–7 years in data centres	Moderate — photonics and communications expertise present	High near-term — data centre interconnect application immediately relevant

Fault-tolerant quantum computing	Exponential for specific problems (factorisation, quantum sim)	NISQ devices only; fault tolerance requires ~10^6 physical qubits	15–25+ years	C-DAC quantum initiative; IIT research groups (quantum hardware limited)	Medium — follow global development; invest in quantum algorithms and software
Adiabatic / reversible CMOS	2–10× for compute-intensive workloads	Research circuit demonstrations; not in production	10–20 years	Academic research only; limited industrial investment	Low near-term — monitor; invest in academic research at IISc and IIT Bombay
3D integrated compute-in-memory	5–20× for memory-bound AI workloads	Early commercial (Samsung HBM-PIM, SK Hynix AiME)	3–8 years, broad deployment	Strong — VLSI design capability at multiple IITs; relevant to India Semiconductor Mission	Very High — near-term, manufacturable, high leverage for AI inference workloads

Maturity assessments reflect the author's synthesis of published research and industry reports as of early 2025. Timelines are inherently uncertain and represent reasonable estimates for widespread commercial deployment rather than proof-of-concept demonstration. CENSE = Centre for Nano Science and Engineering (IISc). C-DAC = Centre for Development of Advanced Computing. Sources: IBM Research (2023); Shastri et al. (2021); National Mission on Quantum Technologies and Applications; India Semiconductor Mission; author synthesis.

10.4 The Circular Economy: Hardware Longevity, Repairability, and Material Recovery

The technological frontiers examined in the preceding sections represent long-term opportunities for transformative improvement in the energy efficiency of computation. The circular economy approaches examined in the present section represent a different kind of opportunity: one that is available now, requires no technological breakthrough, and could deliver significant reductions in the embodied carbon of ICT hardware within years rather than decades, given appropriate policy and market design.

The core principle of the circular economy, as applied to ICT hardware, is to extend the productive life of devices and components for as long as they continue to deliver useful capability, and to ensure that the materials they contain are fully recovered and reused at end of life. This principle is in direct tension with the economic model of the consumer electronics industry — and, to a significant degree, of the enterprise ICT procurement industry — which has been built around regular hardware refresh cycles that generate revenue for manufacturers and distributors but produce significant environmental cost in the form of embodied carbon and e-waste.

The right-to-repair movement — which advocates for the legal right of device owners and independent repair businesses to repair their own equipment, and for manufacturers to provide the spare parts, tools, and documentation needed to enable repair — has gained significant legislative traction in the European Union and in several US states over the past three years. The EU's Ecodesign for Sustainable Products Regulation, which entered into force in 2024, establishes repairability requirements for a range of electronics products and mandates the availability of spare parts for defined periods after product launch. India's Consumer

Protection Act and Bureau of Indian Standards' draft standards for repairability represent the beginnings of a domestic policy framework in this direction, though the specificity and enforceability of these provisions remains limited by comparison with the EU's approach.

For Indian consumers, enterprises, and government bodies, the circular economy principle has an immediate practical application in the form of hardware longevity planning: the deliberate extension of computing equipment life beyond the manufacturer's standard replacement cycle, where the equipment continues to meet performance requirements. A server retained for seven years rather than four years amortises its embodied carbon over a significantly longer operational life, reducing the embodied carbon per unit of computation by a proportional amount. Studies have estimated that extending average smartphone lifetime from 2.5 to 4 years, and average laptop lifetime from 4 to 7 years, would reduce the embodied carbon of consumer electronics by 30 to 40 per cent without any change in manufacturing processes.

It is pertinent to note, in the Indian context, that the Central Government Electronics (CGE) procurement framework — which governs the acquisition of electronics by central government ministries and departments — could be a powerful lever for circular economy principles if its specifications were revised to include minimum durability requirements, repairability indices, and end-of-life take-back obligations for vendors. The volume of electronics procured by the Government of India annually is sufficient to constitute a meaningful market signal if directed toward circular design principles. Similarly, the Government of India's GeM (Government e-Marketplace) portal, which centralises procurement for government entities, could incorporate sustainability scoring for hardware — covering energy efficiency,

repairability, and take-back commitments — into its vendor qualification and product listing criteria.

10.5 Policy Levers: What Governments and Institutions Must Do

The energy challenge of computing cannot be solved by engineers and researchers working in isolation from the policy environment. The efficiency improvements, technological innovations, and circular economy transitions described in the present volume require not only technical ingenuity but also the creation of regulatory conditions, market structures, and institutional capabilities that make energy efficiency more rewarding — and energy waste more costly — than the current baseline. The present section surveys the most important policy levers available to governments and institutions, with particular attention to the instruments available to the Government of India.

Mandatory energy efficiency standards for computing equipment — analogous to the efficiency standards that have successfully reduced the energy consumption of refrigerators, air conditioners, and lighting — represent perhaps the most direct policy intervention available. The Bureau of Energy Efficiency's Star Rating Programme for servers and data centres, currently voluntary, could have substantially greater impact if made mandatory for government procurement and eventually for the broader market. The key to effective standards in this area is the choice of the right metric: a standard based on PUE alone, as noted repeatedly in the present volume, misses the majority of the environmental impact. A standard based on useful work per unit of energy — incorporating server utilisation, efficiency at partial load, and embodied carbon — would be more comprehensive and more gaming-resistant.

Carbon pricing — the imposition of a cost on greenhouse gas emissions, either through a tax or a cap-and-trade mechanism — is the most economically efficient instrument available for internalising the environmental cost of energy consumption. By making carbon emissions costly, carbon pricing creates a continuous incentive for energy efficiency improvement across all sectors, including ICT, without requiring regulators to specify which technologies or practices should be adopted. India does not currently have an economy-wide carbon price, though the Perform, Achieve and Trade (PAT) scheme operated by the Bureau of Energy Efficiency — which caps and trades energy efficiency certificates among large industrial energy consumers — embodies some of the same economic logic. The extension of PAT or a similar scheme to the ICT sector, and its eventual evolution toward a broader carbon pricing mechanism consistent with India's NDC commitments, is a policy direction that the present author regards as both economically sound and environmentally necessary.

Public procurement — the purchasing power of government — is a policy instrument whose potential in the green computing domain is substantially underutilised. The Government of India procures computing equipment, cloud services, and network infrastructure at a scale that, if directed toward sustainability criteria, could constitute a significant market signal. The inclusion of energy efficiency, repairability, circular design, and carbon disclosure requirements in central government procurement specifications — administered through GeM with appropriate technical support from BEE and NIC — would simultaneously reduce the government's own environmental footprint and create commercial incentives for vendors to improve their products' sustainability credentials.

10.6 An Agenda for Indian Research: Open Problems at the Frontier

The preceding chapters of the present volume have identified, at the close of each, specific research directions relevant to the topics of that chapter. It remains, in the closing section of the final chapter, to assemble a broader view of the research agenda that India's universities, research institutions, and national laboratories should, in the author's assessment, prioritise in the domain of green computing over the coming decade.

This agenda is shaped by two considerations. The first is global: the research problems that are most important for the world to solve in order to achieve net-zero computing. The second is distinctively Indian: the research problems for which India's specific capabilities, contexts, and constraints make Indian researchers particularly well-positioned, or particularly obligated, to contribute.

On the global front, the most pressing research priorities — in the author's judgement — are: the development of practical, manufacturable, and programmable in-memory computing architectures that reduce data movement energy for AI workloads; the maturation of neuromorphic and photonic computing from laboratory demonstrations to deployable systems; the development of robust, verifiable methodologies for lifecycle carbon accounting of ICT hardware and software; and the design of regulatory frameworks that make 24/7 carbon-free energy the standard rather than the exception for computing infrastructure.

On the distinctively Indian front, the most important priorities are: the characterisation and improvement of the energy footprint of India's rapidly growing data centre sector, including the development of India-specific baseline data, performance benchmarks, and policy recommendations; the development of

energy-efficient AI systems designed for Indian languages and cultural contexts, so that the AI components of India's digital public infrastructure do not import the energy inefficiency of systems designed for different linguistic and computational contexts; the creation of an India-specific grid carbon intensity API and carbon-aware computing infrastructure that makes temporal and geographic carbon optimisation accessible to Indian cloud users; and the development of research and policy capacity within Indian institutions — including the Bureau of Energy Efficiency, SEBI, MeitY, and the Department of Science and Technology — to evaluate, regulate, and guide the green computing agenda with the technical depth it requires.

The author concludes with a reflection that is, at once, sobering and hopeful. The environmental challenge posed by the growth of computing is real, its trajectory is concerning, and the gap between current practice and what the laws of physics permit is enormous. But that gap is also, precisely, the measure of the opportunity that lies before every engineer, researcher, and policy maker who takes the challenge seriously. The efficiency improvements documented in Chapters 2 through 8, the emerging technologies surveyed in the present chapter, the measurement frameworks described in Chapter 9, and the policy instruments outlined in Section 10.5 are not hypothetical; they are real, they are accessible, and their implementation is a matter of will and priority rather than of fundamental limitation. India's digital infrastructure is, in significant measure, still being built. The choices made now — in the design of data centres, the selection of hardware, the writing of software, the drafting of procurement standards, and the funding of research — will determine the energy profile of that infrastructure for decades. The present volume has been written in the conviction that those choices will be better for having been

made with understanding, and that understanding — of the physics, the engineering, the economics, and the ethics of green computing — is what its chapters have endeavoured to provide.

Table 10.2 Open Research Problems in Green Computing — Priority Assessment for India

Research Problem	Global Priority	India-Specific Relevance	Suggested Lead Institution(s)	Indicative Horizon
Energy-efficient AI inference for Indian language models (Indic NLP at scale)	High	Very High — national AI infrastructure; multilingual models are larger and more energy-intensive per task	IIT Madras (AI4Bharat), IIT Bombay, TCS Research	2–5 years
India grid carbon intensity API and carbon-aware cloud scheduling framework	High	Very High — POSOCO data exists; integration with cloud platforms needed	IIT Delhi, C-DAC, POSOCO, MeitY	2–4 years
Life cycle assessment database for Indian ICT supply chain	High	High — Indian manufacturing under PLI; Scope 3 reporting requires domestic LCA data	CSIR-NEERI, IIT Bombay, Bureau of Indian Standards	3–6 years
3D compute-in-memory architecture for energy-efficient AI hardware	Very High	Very High — aligns with semiconductor mission; IIT VLSI	IIT Bombay, IIT Madras, ISRO Semiconductor Lab,	4–8 years

(India Semiconductor Mission)		expertise applicable	Qualcomm India	
Energy and water audit methodology for Indian data centres (BEE certification)	High	Very High — India-specific climate context; water stress in major DC markets	BEE, IIT Hyderabad, IIT Bombay, Uptime Institute India	2–4 years
Circular economy policy design for Indian electronics (EPR effectiveness)	Medium-High	High — India's informal e-waste sector; PLI hardware manufacturing growth	TERI, CPCB, IIT Delhi, Ministry of Environment	3–5 years
Neuromorphic processor design for IoT and edge AI in Indian agricultural and industrial contexts	Medium	High — BharatNet rural infrastructure; low-power edge computation demand	IISc CENSE, IIT Bombay, C-DAC, agriculture sector partners	5–10 years
Quantum computing energy efficiency and cooling research (NMQTnA)	High globally	Medium-High — NMQTnA investment; India can lead on quantum algorithms and software	C-DAC, IIT Madras, IIT Bombay, IISER Pune	8–15 years

Priority assessments reflect the author's synthesis of global research literature and India's specific technical capabilities, policy landscape, and development priorities. Horizons indicate indicative time to impactful research output, not time to commercial deployment. NMQTnA = National Mission on Quantum Technologies and Applications. PLI = Production Linked Incentive. CSIR-NEERI = CSIR National Environmental Engineering Research Institute. TERI = The Energy and Resources Institute.