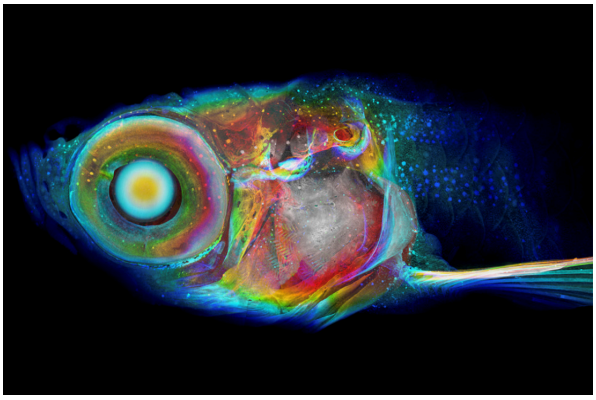
CASC | Coalition for Academic Scientific Computation

Regional Collaboration as Strategic Infrastructure: A Framework for the Future of Research Computing



Summary

The notion that every institution can independently sustain comprehensive Research Computing and Data (RCD) capabilities—an assumption that guided a generation of institutional investment—no longer holds. Amid accelerating pressures, regional RCD collaboration offers the most practical path for institutions to maintain long-term research competitiveness. This paper presents a framework for the collaborations that will shape the future of research computing.

Introduction

Research computing is being reshaped faster than most institutions can respond. Frontier AI training runs can consume, in a few weeks or months, as much compute and energy as many traditional high-performance computing (HPC) projects used over years. Lacking the purchasing scale to compete with hyperscalers, universities find it increasingly difficult to access critical infrastructure and expertise required to support rapidly evolving AI workflows, compliance requirements, and security expectations. At the same time, federal funding faces structural uncertainty, challenging RCD budgets and sustainability [6].

The pressures facing RCD centers today are not the cyclical budget concerns institutions are accustomed to managing. They are structural, they are accelerating, and they make

institutional isolation increasingly costly. Regional collaboration offers the most practical path available at the scale and speed this moment requires. By pooling investments, sharing expertise, and reducing unnecessary duplication, RCD collaborations enable institutions to achieve together what none could accomplish alone, while ensuring that the benefits of advanced research computing reach all corners of the nation's research ecosystem. A 2025 CASC position paper [1] outlined benefits, challenges, and strategies for success in regional RCD collaborations. This paper updates that foundation in light of what has changed since, and offers a practical framework for funders, institutions, and the RCD community.

The Evolving RCD Landscape

Within institutions, increasingly complex scientific workflows and expanding activity in AI and quantum computing have compounded the demands on RCD services. At the same time, external pressures, including funding uncertainties [6], regulatory developments, and heavy AI investments by industry, have complicated RCD procurement cycles, contributed to infrastructure shortcomings, amplified workforce and expertise gaps, and raised new questions about ethical and responsible use of data and AI. Regional collaboration offers a path to transform RCD operations to be more efficient, responsible, and responsive to workforce and societal needs.

What has changed since 2025?

- Rapid uptake of generative AI has driven a supply crisis for GPUs [2] [3], memory, flash storage, and adequately equipped data centers [14][15].
- Shifting federal priorities have altered the outlook for federal funding of research infrastructure [16].
- State-supported public-private models have emerged to accelerate research as a regional economic engine [10][11][12][13].
- Pressures have increased around data centers' carbon footprints and demand for energy [4] and water [8][9].

Science is getting bigger. All fields are becoming more data-intensive, fundamentally altering the infrastructure required to support modern science. GPU-driven computing environments, large-scale storage systems, regulated data enclaves, advanced networking, and specialized software ecosystems are now essential components of research across disciplines. As AI workloads generate unprecedented data volumes, new approaches are required for the storage, curation, and governance of AI training data and models. Quantum computing, a key area for future innovation, also requires new resources and expertise that are likely out of reach for many institutions.

Regional collaboration addresses this scale mismatch through shared resources. A single institution may not be able to justify a regulated data enclave, a quantum testbed, or a petabyte-scale curated data archive for its own research community alone. A regional consortium serving dozens of institutions is more likely to succeed.

RCD infrastructure is becoming harder to maintain. Research infrastructure has become more expensive and difficult to acquire, operate, and refresh. The expansion of AI data centers has fueled a supply crisis, intensifying competition for critical components and leaving institutions struggling to acquire sufficient GPU capacity. At the same time, the rapid evolution of GPU architectures necessitates increasingly frequent refresh cycles. Physical space, power, cooling, and personnel constraints pose additional barriers.

Regional collaboration changes the institution's position in this market. Coordinated procurement gives consortia greater leverage with vendors on both pricing and delivery priority during shortages. Shared facilities concentrate power, cooling, and floor space at sites engineered for advanced infrastructure, rather than requiring every campus to retrofit. Shared refresh planning also allows phasing of hardware investments so that newer capabilities are available somewhere in the region, even when no single institution can afford to refresh on the current cycle.

Funding structures are in flux. Against a backdrop of rising costs, shifts in funding structures are challenging RCD budgets and sustainability.[1] Federal funding has traditionally been a major source of support for RCD infrastructure, but federal grants are limited in scope and duration, and uncertainty surrounding National Science Foundation (NSF) budgets represents a real and urgent risk. Although recent years have seen an increase in state funding for regional public-private collaborations, institutions have also faced enrollment challenges and state funding limitations [5] that may constrain their ability to invest in RCD infrastructure.

RCD leaders are accustomed to managing year-to-year budget variability, but a structural reduction in federal research infrastructure funding would affect RCD operations in ways that incremental belt-tightening cannot absorb. The exposure varies by institution: those whose RCD operations depend heavily on indirect cost recovery from federally funded research, on dedicated federal infrastructure awards (such as NSF Campus Cyberinfrastructure and Major Research Instrumentation awards), or on federally subsidized national resources [7] face larger and more immediate risk than those with diversified funding portfolios. However, diversification is hard to build during a crisis; it must be built before one.

[1] A companion CASC position paper, “Models for Sustainability and Strategic Advancement of Institutional Research Computing and Data,” examines these shifts and presents a framework to inform resilient financial models for RCD.

RCD operations carry high fixed costs in facilities, baseline staffing, and software licensing. Regional collaboration converts a portion of those fixed costs into variable participation costs, which are easier to scale without losing core capability. An institution that contributes to a shared regional facility can retain access even if it cannot increase its contribution; an institution operating alone faces a step function when it can no longer afford the next refresh. In addition, while hardware refresh deferrals during lean times are recoverable, staff departures often are not. RCD professionals who leave during a contraction frequently do not return, and the institutional knowledge they carry is difficult to replace. This makes protecting core human capacity even more critical than protecting hardware investment during a downturn.

Institutions in well-developed regional ecosystems retain access to capabilities, expertise, and continuity that no individual budget could sustain through a major contraction. Institutions that delay collaboration until contraction forces the issue often enter from a weaker negotiating position, with fewer assets to contribute and less leverage in governance discussions.

Specialized expertise is scarce. The infrastructure for AI and quantum computing requires additional expertise beyond traditional HPC. To make advanced technologies usable in research contexts, institutions increasingly require expertise in container orchestration, model serving, MLOps workflows, distributed software stacks, and large-scale data management for AI workflows, as well as algorithm development, workflow integration, and workforce training to support quantum computing. As few institutions possess comprehensive expertise across all these domains, specialized expertise has become a rate-limiting factor, hampering institutions' ability to meet demand.

Regional collaboration enables institutions to share expertise across multiple campuses through distributed support. Shared training programs and cross-institutional mentoring help build the next generation of RCD professionals.

Cybersecurity and compliance are becoming more complex. Institutions face overlapping, and sometimes conflicting, requirements around data protection, access control, and regulatory compliance. Maintaining compliant research environments requires specialized expertise and operational maturity to navigate requirements such as Cybersecurity Maturity Model Certification [17], Federal Information Security Management Act [19] stipulations, National Institute of Standards and Technology guidance [18], state-level data protection laws, and international data agreements. Regional collaboration can help institutions meet these demands through shared security operations centers, coordinated incident response, regulated data enclaves, collective threat intelligence, and harmonized compliance frameworks.

Concerns around responsible AI raise additional considerations. The energy consumption [4], cooling requirements, water usage [9], and electronic waste associated with data centers can complicate institutional commitments to carbon neutrality and environmental sustainability. Regional facilities may improve sustainability by increasing infrastructure utilization, optimizing cooling environments, coordinating energy management, and reducing duplication of underutilized systems. However, not every state has a regional-scale data center available, and successful implementation of new shared data centers will require careful stakeholder engagement.

Why AI Changes the Collaboration Equation

AI is reshaping demand for computing capacity, expanding the expertise required, and introducing new governance questions institutions must answer. AI computing workloads are more uneven and more volatile than traditional HPC workloads; training even a moderately sized model can monopolize an institution's GPU capacity for weeks, while inference workloads spike with little warning. Regional pools can address this challenge by providing larger capacity, which offers more opportunities for efficient scheduling of diverse workloads.

In addition, the AI stack requires expertise that is scarce, not merely expensive. ML engineers, model designers, and researchers who can fine-tune large models are in short supply, especially at salaries academic institutions can offer. Research facilitation is also essential: faculty and students need help translating disciplinary questions into computational workflows, selecting appropriate tools, and using AI systems responsibly. Sharing this expertise through regional collaboration is for many institutions the only realistic path to providing meaningful AI support for faculty and students, while also providing an opportunity for institutions to align on defining the culture governing responsible AI use.

Regions that build collaborative AI capacity now will define how academic AI is governed, resourced, and used over the next decade. Regions that wait will have less influence over the standards, infrastructure, and practices they will ultimately depend on.

The Case for Regional Collaboration

Regional collaboration is not simply a mechanism for reducing costs: it is a strategic framework for accelerating science. Sharing resources creates economies of scale and strengthens resilience in the face of challenging market pressures. RCD collaborations can sustain larger, more capable shared infrastructure; support teams with deeper, specialized expertise; harmonize access, security, and compliance mechanisms; and lower barriers to adopting emerging technologies. Alongside the hardware, and equally as important, regional

collaboration strengthens the people, institutional knowledge, governance structures, and operational practices that make infrastructure usable.

The case for collaboration is often framed in terms of what smaller or less-resourced institutions gain, but well-resourced institutions have equally compelling reasons to participate. Even the largest R1 universities face GPU supply constraints [2], struggle to recruit and retain specialized AI and quantum expertise, and carry growing compliance burdens. Regional collaboration helps institutions diversify risk against federal funding volatility [5], share the cost of regulated enclaves and specialized testbeds, access surge capacity for workloads that exceed local resources, and shape policies on responsible AI use, data stewardship, and environmental accountability. These benefits come with real tradeoffs—shared governance is slower than unilateral decision-making, and shared infrastructure requires accepting constraints that purely institutional resources do not impose. However, the alternative of maintaining institutional self-sufficiency at the scale modern research now requires is increasingly out of reach even for the best-resourced universities.

Collaboration also creates access pathways for institutions that otherwise could not participate meaningfully in advanced computational research, including emerging research institutions, minority-serving institutions, community colleges, rural campuses, and institutions in under-resourced regions. However, collaboration should not require every participant to become a replica of a large R1 university, nor should it relegate smaller partners to act as consumers of services they did not help design, governed by priorities they did not help set. That is not collaboration; it is hosting. Successful ecosystems recognize and amplify the distinct strengths of diverse institutions, offering smaller institutions pathways into operational and governance roles. In this way, collaboration is a strategy through which institutions of all sizes hedge against shared risks and build capabilities that no single institution can fully sustain alone.

A Framework for Regional Collaboration

Effective and enduring collaborations consistently share four foundational characteristics:

1. **A sustained catalyst.** Drivers such as funding opportunities, institutional champions, strategic crises, or state-level investments are critical to overcoming the structural and cultural hurdles that impede collaborations. The catalyst may differ by region, but it must be sustained long enough to move from idea to action.
2. **Institutional buy-in.** All partners must agree that the benefits of collaboration are worth the investments and operational complexity it requires. Aligning shared goals with each institution's mission is essential to achieving and sustaining buy-in.

3. **Sustained funding.** Collaborations require sustainable financial models capable of supporting operations beyond initial infrastructure deployment. In addition to investing in shared resources, funding the connective systems to facilitate, coordinate, and support inter-institutional collaboration is critical.
4. **Trusted governance.** Governance operationalizes institutional alignment with mechanisms for ensuring accountability to shared goals and processes. Governance structures should be transparent, equitable, and resilient to leadership transitions.

Regional Collaboration Models

Collaboration should not force partners into the same mold; rather, a contribution framework should identify and amplify what each institution brings to the broader research ecosystem and facilitate connectivity to regional resources. To support participation across the full spectrum of institutional size, type, and resource level, collaboration models should enable participants from all involved institutions to make meaningful, mission-relevant contributions from the conceptualization phase through all aspects of the collaboration's operations and governance.

A wide range of collaboration models can be successful, and no single model fits every region. Appendix A presents several models that have proven to be feasible and effective for sustained inter-institutional collaboration.

Shared Services Beyond Compute

Some of the highest-value collaborative opportunities emerge not from shared hardware, but from shared operational services. These services reduce duplication, expand access to specialized expertise, and strengthen the usability and adoption of regional infrastructure.

Shared Service Area	Value to Institutions
Storage and archives	Lower storage costs, better data stewardship, and centralized governance
Software licensing	Reduced duplication and access to a broader software portfolio
Research facilitation expertise	Expanded support capacity in domain-specific areas
Cybersecurity operations	Improved compliance and incident response
Training programs	Workforce development, onboarding, adoption of shared services

Financial Models

Short-term grant funding can catalyze a regional ecosystem but rarely sustains one. Collaborations require long-term financial commitments, and most combine multiple funding

approaches, including institutional contributions, grant support, state or regional investment, philanthropic gifts, service fees, and in-kind commitments. Appendix A summarizes the strengths and risks of different financial structures for regional collaborations.

In addition to monetary contributions, it is important to explicitly recognize the value of non-monetary contributions such as staffing, software licenses, networking infrastructure, training programs, facilities, and operational expertise. Recognizing these contributions helps broaden participation, especially for institutions that may not be able to contribute equally in cash but can still strengthen the collaboration in meaningful ways.

Governance as Core Infrastructure

Governance is often the determining factor in whether collaborations succeed or fail. Cultural misalignment, conflicting priorities or strategies, differing budgets, and incompatible legal or policy frameworks across institutions can derail collaborations. Shared governance helps establish trust by clarifying how decisions will be made, how institutions will be represented, and how accountability will be maintained. With the right governance structure in place, contributors are better positioned to communicate effectively across institutions, make decisions that account for different needs and constraints, and adapt to changes and challenges as they emerge.

Collaboration adds complexity; building consensus requires more time, discussion, and processes when a broader range of stakeholders is involved. Sharing environments can also introduce risks and liability considerations; for example, if a breach occurs at a shared facility, where does the responsibility lie? Governance is crucial to answering such questions and navigating the inevitable challenges.

Effective governance structures should:

- provide meaningful representation for all participating institutions,
- clearly define roles and include decision-makers with institutional authority,
- maintain small decision-making bodies with transparent communication practices,
- define conflict-resolution mechanisms, and
- support long-term continuity despite leadership turnover.

Measuring Success

The success of regional collaboration should not be measured solely by infrastructure scale. Effective ecosystems create broader scientific, operational, workforce, and societal value. Success measures should also capture the capabilities, relationships, and resilience that the collaboration builds over time.

Category	Indicators
Access and Participation	Increased participation of emerging research institutions and minority-serving institutions, broader geographic access, participation across institution types
Operational Effectiveness	Higher utilization, reduced duplication, faster onboarding
Workforce Development	Growth in trained users and RSE personnel, shared staffing capacity, and cross-institutional mentoring
Research Outcomes	Publications, grants, and cross-institutional collaborations
Sustainability	Reduced energy usage, diversified funding, and long-term financial sustainability
Governance and Trust	Sustained institutional participation, transparent decision-making, continuity through leadership transitions

Conclusions & Recommendations

Regional collaboration is no longer an optional enhancement to institutional RCD strategy; it is essential for maintaining long-term research competitiveness. The scale, complexity, and cost of modern computational research now exceed what most institutions can independently sustain. Institutions that recognize this and act early will shape the regional ecosystems of the next decade. Institutions that delay will inherit arrangements designed without them, on terms they did not set.

Federal funding uncertainty, GPU supply constraints, and the rapid consolidation of AI capabilities outside academia are not temporary disruptions—they are structural shifts that reward coordinated action and make institutional isolation costly. Regions that build collaborative capacity now will define how academic research computing is governed, funded, and accessed for the next generation of science. Regions that wait will have less influence over the standards, infrastructure, and governance models they will depend on.

Collaboration is not fundamentally a technical challenge. It is organizational, cultural, financial, and political. Sustainable collaboration requires trust, transparency and long-term investment. The ecosystem functions best when every institution contributes something meaningful, regardless of size or resource level. The institutions that step forward to lead will determine whether the nation's research computing future is coherent or fragmented, contribution-focused or stratified, resilient or fragile.

Recommendations for Funding Agencies
Funding agencies should expand incentives to encourage regional approaches in infrastructure programs. Funding individual institutions to build duplicative capabilities is no longer a defensible use of constrained federal resources. Agencies should also fund the connective tissue that determines whether collaborations succeed or collapse: governance coordination, interoperability, workforce development, and operational support for under-resourced institutions.

Recommendations for State Governments
State governments should recognize that regional RCD infrastructure is economic development infrastructure. States that invest in coordinated AI and computing capacity will see returns in workforce development, industry partnerships, and research-driven economic growth. States that leave this to individual institutions will find their research universities competing with each other for resources that would be more productively pooled. AI and RCD initiatives should be approached as opportunities to strengthen workforce development, economic competitiveness, and research capacity across regional ecosystems.

Recommendations for Institutional Leaders
Institutional leaders should treat regional collaboration as a core infrastructure strategy, not a supplemental partnership activity. This means committing to multi-year funding, assigning accountability to senior leadership, and accepting governance arrangements that constrain unilateral institutional action in exchange for collective capability. Long-term investments in governance, workforce development, and shared operational capacity are as important as hardware acquisition.

Recommendations for RCD Leaders
RCD leaders should resist the temptation to defer collaboration until conditions are easier. Lightweight coordination mechanisms should be established early, before they are forced by a crisis. Shared workforce development deserves investment alongside infrastructure planning, and non-financial institutional contributions—staffing, expertise, software, networking—should be explicitly recognized and valued alongside monetary contributions.

Authors: *Katia Bulekova (Boston University), Suranga Edirisinghe (Georgia State University), Tabitha Samuel (University of Tennessee), Bruno Abreu (Pittsburgh Supercomputing Center), Christy Long (Oregon State University), Mike Navicky (University of Virginia), Ruth Marinshaw (Yale University), Kim Wong (University of Pittsburgh), Jaime Combariza (University of Pennsylvania), Jason Simms (Swarthmore College), Jill Gemmill (Clemson University), Carolyn Casler (CASC), Kathryn Kelley (CASC)*
With the support of Anne Johnson, Creative Science Writing

About CASC

The Coalition for Academic Scientific Computation is an educational nonprofit 501(c)(3) organization with 110+ member institutions representing many of the nation's most forward-thinking universities and computing centers. CASC is dedicated to advocating for the use of the most advanced computing technology to accelerate scientific discovery for national competitiveness, global security, and economic success, as well as develop a diverse and well-prepared 21st century workforce. Learn more at http://casc.org.

References

[1] Coalition for Academic Scientific Computation. (2025). Teaming Up for Impact: Regional Collaborations for Research Computing and Data. https://casc.org/policy-publications/position-paper/teaming-up-for-impact-regional-collaborations-for-research-computing-and-data/.

[2] Kudiabor, H. (November 20, 2024). AI's Computing Gap: Academics Lack Access to Powerful Chips Needed for Research. Nature News. doi:10.1038/D41586-024-03792-6.

[3] Morgan, K. and Partridge, B. (December 2, 2025). AI's global resource race. S&P Global. https://www.spglobal.com/en/research-insights/special-reports/look-forward/data-center-frontiers/global-ai-power-demand-challenges-opportunities.

[4] Shehabi, A., et. al. (December 2024). 2024 United States Data Center Energy Usage Report. Lawrence Berkeley National Laboratory. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report_1.pdf.

[5] Darraro, L. (January 6, 2026). Uncertainty Looms Large for US Science Funding in 2026. C&EN. https://cen.acs.org/policy/research-funding/Uncertainty-looms-large-US-science/104/web/2026/01.

[6] Computing Research Association. (April 24, 2026). NSF FY 2027 Request: Another Potentially Disastrous Budget Request, with Proposed Deep Cuts for Essential Areas of Computing Research. Computing Research Policy Blog. https://cra.org/govaffairs/blog/2026/04/nsf-fy2027-pbr/.

[7] NSF National Artificial Intelligence Research Resource. https://www.nsf.gov/focus-areas/ai/nairr.

[8] Privette, A.P., Barros, A., and Cai, X. (2026). Data Centers Water Footprint: The Need for More Transparency. *AGU Advances*, 7(2), doi:10.1029/2025AV002140.

[9] Yañez-Barnuevo, M. (June 25, 2025). Data Centers and Water Consumption. Environmental and Energy Study Institute. https://www.eesi.org/articles/view/data-centers-and-water-consumption

[10] Office of Governor Kathy Hochul. (January 30, 2026). Governor Hochul Announces Empire AI SUNY Campus Partnerships to Expand Access to Artificial Intelligence Use for the Public Good. https://www.governor.ny.gov/news/governor-hochul-announces-empire-ai-suny-campus-partnerships-expand-access-artificial.

[11] Salazar, J. (October 8, 2025). A Bright Future in Texas Computational Science. TACC. https://tacc.utexas.edu/news/latest-news/2025/10/08/a-bright-future-in-texas-computational-science/.

[12] Keystone AI + Quantum Factory. https://keystonefactory.org.

[13] Office of Governor Maura Healey. (May 6, 2025). Governor Healey Advances State's AI Leadership with Major Investments in Massachusetts AI Hub. https://www.mass.gov/news/governor-healey-advances-states-ai-leadership-with-major-investments-in-massachusetts-ai-hub.
[14] United Nations University – Institute for Water, Environment and Health (UNU-INWEH). (June 3, 2026). Environmental Cost of AI's Energy Use: Carbon, Water and Land Footprints. https://unu.edu/inweh/news/environmental-cost-of-AIs-Enrgy-use-carbon-water-and-land-footprints
[15] de Vries-Gao, A. (2026). The Carbon and Water Footprints of Data Centers and What This Could Mean for Artificial Intelligence. *Patterns*, 7(1), 101430, doi:10.1016/j.patter.2025.101430.
[16] Agarwal, K. (May 1, 2026). House Bill Cuts Funding for NSF, NASA Science. Association of American Universities. https://www.aau.edu/newsroom/leading-research-universities-report/house-bill-cuts-funding-nsf-nasa-science.
[17] U.S. Department of Defense, Office of the CIO. Cybersecurity Maturity Model Certification (CMMC) Model Overview. https://dodcio.defense.gov/Portals/0/Documents/CMMC/ModelOverview.pdf
[18] National Institute of Standards and Technology (NIST). Special Publication 800-171: Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. NIST Computer Security Resource Center. https://csrc.nist.gov/pubs/sp/800/171/r3/final.
[19 Cybersecurity and Infrastructure Security Agency. Federal Information Security Modernization Act of 2014 (FISMA). https://www.cisa.gov/topics/cyber-threats-and-advisories/federal-information-security-modernization-act.

Appendix A: Regional Collaboration Models and Funding Structures

Generalized Regional Collaboration Models

Model	Description	Advantages	Challenges	Best Fit	Example
Shared Facility	Jointly operated physical infrastructure	Economies of scale, centralized operations	High upfront capital costs	Regions with power/space constraints	Massachusetts Green High Performance Computing Center (MGHPCC) (https://mghpcc.org)
Centralized Shared System	Consortium-operated shared compute system	Lower barriers to participation	Allocation governance complexity	Mid-sized consortia	Rocky Mountain Advanced Computing Center (RMACC Alpine) *(https://www.colorado.edu/partnerships/rmacc/)*
Federated Resource Model	Institutions share portions of local resources	Preserves institutional autonomy	Interoperability challenges	*Existing RCD ecosystems*	OneOklahoma Cyberinfrastructure Initiative (OneOCII*) (https://www.oneociiokepscor.org/)*
Public-private partnership	Public-private-supported AI/RCD infrastructure	Large-scale investment and equity potential	Political and funding dependency	States prioritizing AI competitiveness	Empire AI (New York) (https://www.empireai.edu)

Generalized Funding Structures

Funding Model	Characteristics	Strengths	Risks
Contribution-Based	Institutions contribute based on capacity	Flexible and inclusive	Potential imbalance in participation
Service-Access	Usage-based pricing	Sustainable cost recovery	Access barriers if pricing is too high
State-Funded	Public investment supports infrastructure	Large-scale deployment potential	Political dependence
Hybrid	Combines multiple models	Diversified and resilient	More complex governance