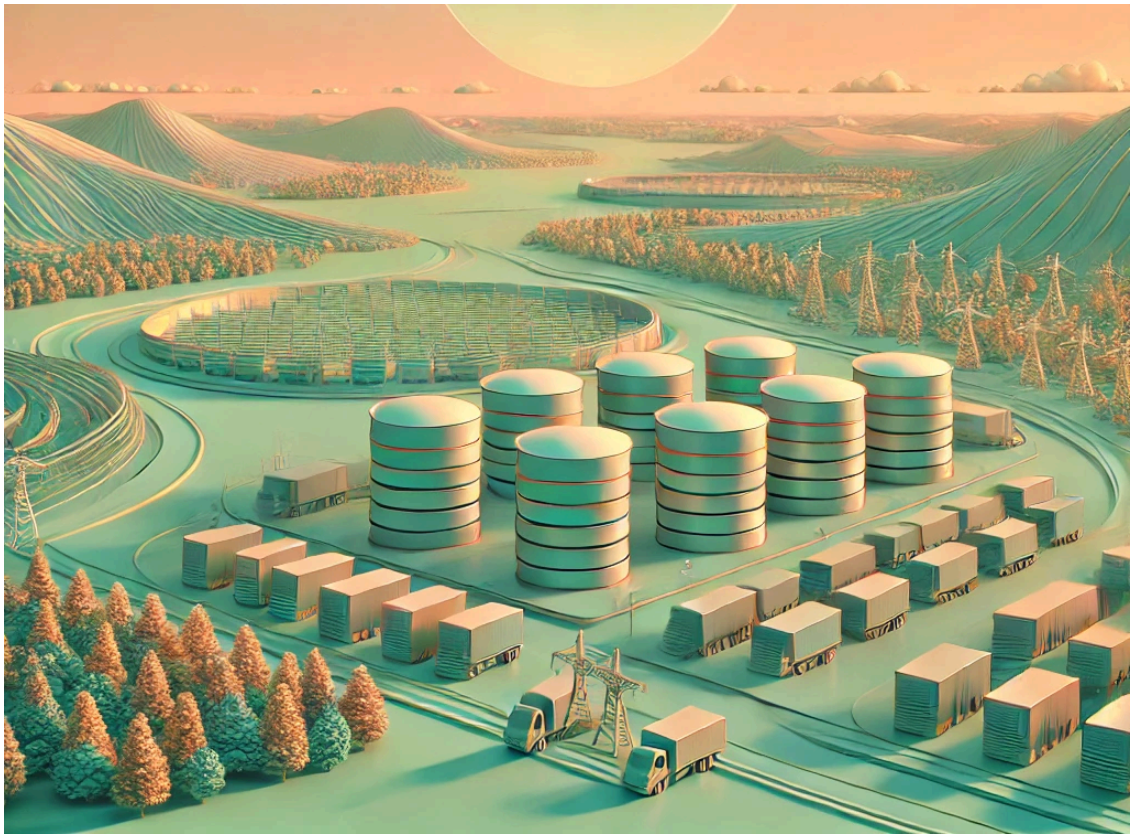
Micro is the new mega

A note on micro data centres prepared by people+ai, EkStep Foundation

Authors:
Shreya Mandi, Srinivas Varadarajan, Kalyan Mangalapalli, Tanvi Lall, Mohan Srinivas, Sasank Chilamkurthy, Swaroop Rajagopalan

September 2024



We are very thankful to our reviewers for their help and guidance: Prasanna Kakhandaki, Manu Awasthi, Divyansh Goel, Ignesius Thambyraj, Anand Hariharan, Ameya Pimpalgaonkar.

India is on the path to becoming a global leader in AI, but realising this vision requires a robust and scalable infrastructure. A distributed network of micro data centres (MDCs), designed to handle critical CPU and GPU workloads while occupying significantly less space and demanding lower upfront investments, represents the future of accessible, scalable, and cost-effective AI infrastructure in India.

This paper delves into the key components and requirements for establishing micro data centres, drawing on our research to define their scope and functionality. We categorise a micro data centre as one with a capacity of 25-300 kW, typically occupying around 800 to 3000 square feet. These data centres are crucial for expanding AI capabilities to the edge, enabling sustainable development through the integration of renewable energy. The flexible CPU-GPU ratio allows MDCs to scale efficiently, providing resilient and self-sufficient compute power necessary for India's growing AI use cases.

Currently, India hosts fewer than 10 micro data centres. Although MDCs lack a standardised definition or size, they are emerging as a critical segment in the hosting infrastructure landscape. Existing MDCs in India serve both domestic and international clients, but there is a pressing need to expand beyond mega data centres to include more micro facilities. The paper explores their use cases in sectors such as healthcare, banking, financial services, insurance (BFSI), and large-scale government operations. The increasing demand at the edge, driven by population growth in Tier II and Tier III cities and the rise of engineering universities focused on deep learning, highlights the importance of these centres.

The paper also examines the capital expenditure (CAPEX) and operational expenditure (OPEX) models associated with micro data centres. Our analysis suggests that building green MDCs can enhance cost-effectiveness, providing a compelling economic model. We predict that an investment of INR 60 crores in a MDC could yield a return of up to 3 times that of a larger data centre. Financing strategies and the potential impact on overall economic growth are also discussed.

Finally, we review the current policies in India that support the development and deployment of micro data centres. Major costs associated with software licenses and certifications (e.g., Uptime and TIA-942) need to be addressed through standardization and improved policy frameworks. Government initiatives, similar to the Udaan scheme, are necessary to foster the growth of smaller players in the MDC market.

This endeavour requires collaboration across various stakeholders, including compute users in both the private and public sectors, government ministries and states responsible for policy creation and enforcement, and investors. The concept of Open Cloud Compute suggests that if hundreds of smaller players can operate collectively like a large cloud provider, a network of micro players can function like a mega network, driving the next phase of AI infrastructure development in India.

India’s data centre landscape

India’s data centre supply is rapidly growing

As India's appetite for data-driven services grows, its data centres are rapidly becoming major energy consumers. A good way to assess the growth of data centres in India is by expressing the capacity in terms of power usage. India’s data centre capacity was projected to be at 1,048 MW at the end of 2023 and reach 1,800 MW by end of 2025.[1]

In terms of their tier classifications for reliability, Tier III[2,3] data centres have dominated the Indian market thus far and are expected to grow to 3365 MW capacity by 2029. Similarly, Tier IV data centres operated at a capacity of 212 MW in 2021 and is expected to increase to 1380 MW by 2029 at a CAGR of 29.54%.[4]

In terms of India’s computing requirements for the coming decade, recommendations from an expert committee set up by the Ministry of Electronics and Information Technology (MeitY) in 2023 recommended that India's national capacity be increased to *"3,000 Petaflops (PF) constituting an inference farm of 2,500 PF and edge compute system of 500 PF”*.[5]

---

[1] CBRE report
[2] Tier III corresponds to 99.6% uptime, approximately 28.8 hours of downtime per year
[3] Data Centre Tier Certification - Uptime Institute
[4] India Data Centre Market Size & Share Analysis - Industry Research Report - Growth Trends
[5] India AI programme entailing over Rs 10000 cr investments to be taken up under PPP mode: MoS IT, ET Government

Large data centres in Tier I cities dominate the market

There is a diverse set of players who act as data centre and compute providers in India. The market is dominated by a mix of Indian and multinational companies. Organisations play to their legacy strengths, invest in long-term partnerships with their suppliers and try to standardise operating models that will be cost-efficient.

Data centre players in India can be categorised into five categories as listed below:

- **Category 1:** Infrastructure players like AdaniConneX (Adani Group), and Yotta Infrastructure (Hiranandani Group).
- **Category 2**: Pure-play data centre companies like NTT, Nxtra by Airtel and Colt DCS who specialise in building out data centres and IT infrastructure.
- **Category 3**: Indian cloud services providers like E2E Networks, Jarvislabs.ai and Vigyanlabs who either focus on the software stack and leverage the data centres of category 1 and category 2 players listed above or offer a vertically integrated solution of hardware and software in their own micro data centre.
- **Category 4**: Hyperscalers like Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP) and Oracle Cloud Infrastructure (OCI) who have a vertically integrated solution.
- **Category 5**: Chip manufacturers' like AMD, Intel and NVIDIA who operate through "Cloud Partnerships" with Yotta and E2E.

The following players (in no particular order) make up nearly **40% of the market**[6]:

- Equinix Inc.
- NTT Ltd.
- Nxtra Data Ltd.
- Sify Technologies Ltd.
- STT GDC Pte Ltd.

The remainder of the market is fragmented across local players or other global service providers.

Mumbai, Chennai, Bangalore and Delhi-NCR are the established data centre markets in India accounting for about **87%** of the country's data centre stock **across ~880 MW and ~13 million sq ft** as of 2023.[7]

- Mumbai- 52%
- Chennai- 15%
- Bangalore- 10%
- Delhi- 10%
- Hyderabad- 5%
- Pune- 7%
- Kolkata- 1%

[6] India Data Centre Companies
[7] CBRE report

MeitY empanelment of data centres

MeitY has built a cloud to accelerate delivery of e-services in the country while optimising Information and Communication Technology (ICT) spending of the Government. To be a part of this cloud, a data centre provider must be empanelled. Empanelment requires a data centre to fall into one of the categories defined as Basic Cloud Services, Advanced Cloud Services or Managed Services. The empanelment primarily pertains to meeting compute, storage, networking and security requirements that are established by MeitY.[8]

All data centres recognised by MeitY for empanelment today are 1 MW plus mega data centres. Today, there are many government workloads, including those of use cases like Aadhaar that require a data centre of less than 1 MW capacity. Micro data centres with sub-1 MW capacity could be utilised for such use cases if they are supported by the government. More about MeitY empanelment and the investor profiles of empanelled data centres in India here.

Data centres are moving towards edge and distributed computing

Globally and in India companies are hedging bets that edge and distributed computing are the future.[9]

For instance Oracle's latest cloud innovations highlight a significant shift towards smaller, more distributed computing environments and edge computing.[10] Key points include:

1. **Distributed cloud approach**: Enables deployment of AI and cloud services across various locations, including edge environments.

2. **Enhanced edge computing**: Introduction of the OCI Roving Edge Device with AI-optimised GPUs for remote processing.

3. **Compact cloud solutions**: The new OCI Dedicated Region25 offers a smaller, scalable setup for on-premises mini cloud environments.

4. **Focus on data sovereignty**: Partnerships to allow organisations to build and maintain AI models within their own countries.

5. **Multi-cloud flexibility**: Collaborations with major cloud providers to offer seamless integration of services across different platforms.

These developments underscore a trend towards more flexible, localised, and edge-oriented computing solutions, addressing the growing demand for processing power closer to data sources and users. This shift aims to meet diverse needs such as low latency, data privacy, and AI processing in remote or disconnected locations.

---

[8] GI Cloud (MeghRaj) | Ministry of Electronics and Information Technology, Government of India
[9] J.P. Morgan makes strategic investment in Edge Laboratories and Evooq
[10] Oracle Expands its Distributed Cloud Capabilities to Help Organizations Innovate with AI

More on research and development in edge computing here

Overview of micro data centres

This section covers the following aspects of a micro data centre:

- Definition of a MDC
- Can MDCs be green?
- Benefits and challenges of being a MDC-strong nation
- Recognising MDCs as a reliable class of compute

Defining a micro data centre (MDC)

Our proposal is that micro data centres should be defined as data centres that:

1. Require power capacity between 25 kW to 300 kW.
2. Occupy around 800 to 3000 square feet
3. Able to accommodate both CPU and GPU-intensive workloads
4. Require less upfront capital investment
5. Located in Tier II, III and IV cities in India, closer to the end user
6. Use renewable energy sources as primary or secondary power source and has lower reliance on main power grid
7. Offers resilience by relying on many small centres instead of one mega data centre

We see two form factors when it comes to micro data centres in India:

- **Traditional MDCs** where all subsystems are integrated and organised in racks on a fixed floor plan.
  Example: Vigyanlabs, Jarvislabs.ai; they have the components of a typical data centre:
  - ➔ Server racks/enclosures
  - ➔ High-performance servers
  - ➔ Networking equipment
  - ➔ Power Distribution Units (PDUs)
  - ➔ Uninterruptible Power Supplies (UPSes)
  - ➔ In-rack/in-container cooling
  - ➔ Environmental monitoring
  - ➔ Solar Panels
  - ➔ Battery Storage
  - ➔ Physical security measures
  - ➔ Fire suppression technology
  - ➔ Diesel generators
  - ➔ Datacentre infrastructure management (DCIM) controller

- **Containerized MDCs** that have flexible cooling and power needs and can be quickly and easily deployed to any chosen location.

Example: PlanckDot, Delta Energy Systems; In addition to the traditional data centre requirements, they might require:

- ➔ In-built battery
- ➔ Condensers

Understanding specifics of a few players operating MDCs in India

There are 10 functional MDCs in India.

Smaller Service or Infrastructure Providers	Location and capacity	Data centre operations and customers	Funding
E2E	Collocated with Netmagic's data centre in Noida and Mumbai	Startups and enterprises in e-commerce, BFSI, AI/ML startups	Public listed company
NetForChoice Solutions	Noida and Bangalore	Colocation and cloud connectivity	No venture funding raised
RackBank Datacentres Pvt. Ltd	Indore and Gurugram; 1MW	Managed and unmanaged dedicated server hosting services	Venture backed
Assistanz Networks	Chennai	Server management and cloud solutions, supporting data centres	Possibly venture backing
Vigyanlabs	Mysore	BFSI, Instant messaging providers, healthcare, e-commerce, Telecom	Self Funded
Jarvislabs.ai	Coimbatore	Foreign B2C, Kagglers	
PlanckDot	Containerised data centres	Small startups and businesses	Venture backed

Discoverability of these providers is difficult unless you are looking for them. The larger CSPs tend to be the natural choice for compute and not the MDCs.

Insight from providers:

- MDCs have the potential to be greener and serve more AI compute needs.
- However some road blocks such as empanelment and certifications can hinder them from being chosen for the user's compute needs.
- Visible here is a trend leaning towards establishing more DCs in Tier II and III cities and in establishing greener technologies, specially in MDCs.

	Micro Data Centres		
	Vigyan Labs	**jarvislabs.ai**	**PlanckDot**
Capacity	50 Racks, 300 kW	Operating out of other data centres' infrastructure (initially started with own data centre)	>5 Racks, 12 kW per rack (PlankDot does containerised setups for other users, and does not own/rent data centres)

Location	Mysore	Coimbatore	Shippable and Containerised (Based in Delhi)
Services and Compute offerings	Mostly CPU and some GPU	Mostly GPU- as a notebook service	CPU
Customers	Instant messaging services, BFSI, Healthcare, E-Commerce, Telecom	Foreign service companies, Kagglers(Global B2C)	Small business
MeitY Empanelment	Nil	Nil	Nil
USP	Sustainable and Green DCs	An AI centric compute offering	Containerised and shippable solution with no setup time and a reduced carbon footprint.

Can a MDC be green?

The answer is YES. Micro data centres can lead the shift towards sustainability in India. There are MDC models that exist today in India that rely solely on renewable energy.

Data centres account for 2% of India's power consumption with the power consumption reaching nearly 140 billion kWh in June, 2023. Most of India's energy comes from coal which poses a huge challenge for the data centre industry with power consumption growing at 4.4% year after year.

Many data centre companies have started to move to renewable sources. Nxtra had declared it will source 140 GWh of renewable energy for its data centres through partnerships with solar and wind projects. AdaniConnex has raised funding to build an environmentally conscious 1 GW data centre. STT Global Data Centres has achieved its target of sourcing 60% of its energy through renewable energy in 2023.[11]

There is a power crisis looming in an AI-powered future. Every AI search can consume as much as 10 times power as normal Google search consumes. Training an LLM can take as much power as that of 5 cars during their lifetime. Google and Microsoft have reported that their greenhouse gas emissions have risen by 50 and 30% respectively from 2019 to 2023. All of this is conducive to the fact that AI is power hungry and the advent of larger models to train and inference is making the carbon footprint steeper. Apart from running the servers the cooling of GPUs consumes significant power.

With this in mind, switching to renewable sources to power AI would mean lower carbon footprints of AI applications and could be key to building sustainable compute infrastructure for AI. **Micro data centres can lead this shift being built with sustainability in mind and reaching near 100% renewable energy consumption.** Vigyanlabs is one such example. They are a MDC powered by rooftop solar energy - 100%.

[11] https://www.mercomindia.com/data-centres-drive-demand-for-renewable-energy

There are many design and operational optimisations that could make a data centre green. Listed are some of the changes that have been made at Vigyanlabs:

- zero water usage for data centre cooling
- fully solar powered, grid independence
- green building design
- high-efficiency design
- AI-driven data centre automation
- AI-optimized energy usage
- AI-enhanced infrastructure management
- use of open-source and custom-built tools, zero CO_2 emissions, Tier-3/TIA-942 compliance

The adoption of AI technology in automation and management along with the shift to renewable energy sources has made it easier to create a 'Green' MDC which is reliable.

Refer to the appendix section for details on designing and building a green MDC.

Benefits of a MDC in the Indian context

Operating requirements of an MDC are smaller in size and allow more flexibility

1. Previously MDCs were not cost effective for either expenses related to capital and operations. However, new approaches to building green data centres make the MDCs highly cost effective in terms of Capex as well as Opex. An example of such a Green Micro Data Centre is the Vigyanlabs data centre in Mysore.
2. Micro data centres due to their smaller size and geographic distribution can be more affordable on multiple fronts:
    a. Land in Tier II, III and IV cities is cheaper than land available in Tier I cities.
    b. Finding large land parcels in densely populated markets is not easy; moving workloads to many smaller centres as opposed to relying on one large centre can alleviate this concern.
    c. Electricity costs for a MDC due to its smaller size will be lower than that of a mega DC.
    d. MDCs of a certain capacity (and below) can be wholly powered by renewable energy sources like solar and wind; thereby reducing the dependence on the grid and making the DC green.
    e. Labour costs and certain state-given tax incentives further reduce operating costs

Micro in power consumptions

Data centres are power hungry.[12] MDCs consume a mere 25 to 300 kW in comparison to even retail stores today that can be 12 kW to malls reaching many MW. Micro data centres

[12] https://timesofindia.indiatimes.com/technology/tech-news/google-has-been-denied-permission-to-build-another-data-centre-in-ireland-heres-why/articleshow/112846737.cms

are generally more capable of maintaining a lower PUE[13] than mega data centres due to their smaller size, simpler infrastructure, and greater operational flexibility (a lower PUE means the data centre is more energy-efficient, using less energy for things like cooling and more for actual computing).

MDCs address the AI power crisis by using green energy sources

GPUs are power hungry and heat up. Apart from running the servers the cooling consumes power as well. Building sustainably can allow AI to keep growing at its current pace. MDCs will house these GPUs and begin a league above by using renewable energy instead of having to transition in the future.
*Additionally, There have been studies about strategic locations of the DCs in which the locations are mapped with weather profiles. Some leading data centre providers like Facebook and Google have been building data centres in icy regions.[14] In regions where nights are much cooler than the day some of the bigger players are using this temperature variation to cool the water which can then be used during the day for cooling.

Resilient to power outages of major cities

Power consumption in major cities like Mumbai is at an all-time high and energy infrastructure is being stressed. Relying on large data centres, located in a few centralised hubs does not make for a resilient hosting infrastructure.[15] The power infrastructure in India today can meet data centre power requirements not only in the Tier I cities but also in many Tier II and III cities where MDCs can be built.[16]
*While Mumbai offers superior connectivity through underwater cables and hence hosts many of the large DCs, other such locations can be identified.

MDCs can provide distributed edge compute by maintaining low latency advantages

Our vision of MDCs spread across the country, creating a distributed compute network. AI solutions deployed on MDCs will have lower inference times on account of being closer to the users, allowing better user experiences and greater adoption of previously challenging solutions. In this paper[17] measures of latency show that while 58% of end-users can reach a nearby edge server in less than 10 milliseconds, only 29% of end-users obtain a similar latency from a nearby cloud location.

Many MDCs operating across the country provide reliability and scalability

Large data centres are a single point of failure, with multiple smaller data centres the risks associated with failure are not as wide spread. Thus a decentralised network of micro data centres provides resilience.

Taking MDCs to Tier II and Tier III cities creates employment opportunities and skilled employable workforce

Data centres require skilled operators, managers and engineers to build and operate. The introduction of such curriculum into the universities and colleges will create a job ready

[13]PUE is a key metric used to evaluate the energy efficiency of data centres. It is calculated by dividing the total amount of power consumed by a data centre by the power used specifically for its IT equipment.
[14] Is it a good idea to build a data centre in cold regions? | Webwerks.
[15] ToI on high power consumptions: Mumbai: Power consumption rises to 3825 MW, crisis on cards
[16]Mumbai News: City's Power Consumption At All-Time High On June 9
[17] Latency Comparison of Cloud Datacenters and Edge Servers

workforce and the establishment of the data centres in lower tiered cities will create employment opportunities for many Indians.
Read more about human resources for MDCs in this appendix section.

Macro challenges faced by MDCs in India

Competing with benefits of recognition that mega data centres avail

Mega data centres enjoy the benefits of recognition as assets to economic development. This enables establishments like STPI (Software Technology Parks of India) and the support of Special Economic Zones (SEZs).

These promote data centres by:

- quality infrastructure
- fiscal package benefits
- with the minimum possible regulations both at the Centre and the State level

By recognising and enabling the same for MDCs, scaling economics can be factored out and allow MDCs to stand as a market alongside mega data centres. Conventional approaches to Building MDCs require high capital expenditure upfront which is solved for with the green and cost effective approach to building them.

Cost of capital is high as there is no recognised bracket for micro data centres

MDCs are seen as high risk investments, making it harder for them to acquire capital. The interest rates on the debt is high; recognition and subsidy by government intervention would make capital investments feasible to MDCs.

Infrastructure, network supply costs and accessibility are hindering

Most Tier II cities have 1 Gbps, 10 Gbps backbone fibres of the Telcos (Airtel, Jio, TCL, Railtel, Vodafone, PowerGrid, BSNL). Tier-III cities have smaller capacity links. Many Tier II and Tier III cities do not have reliable and high speed network infrastructure which is essential for compute services. Bringing fibre optics to places is expensive.

Have to comply with expensive international standards

In India today, there exists no established criteria by which MDCs are recognised. Consequently, MDCs rely on compliance and safety standards and requirements that have been established for much larger centres; typically resulting in significant operational expenditure to be recognized as reliable compute providers and a hindrance when investors and operators view the viability of building a micro data centre. On the other hand, not every device that has computational capacity qualifies to be an MDC, and classifying them as micro data centres might hamper the real value of a reliable MDC. It can be as expensive as $200,000 for a MDC to be compliant to TCDD, TCCF and TCOS certifications from UPTIME and becomes infeasible for them to maintain compliance.

The government can define clear standards and requirements for power, safety and reliability metrics in the Indian context that a MDC needs to meet.

Disparity in visibility

Currently only data centres over 1 MW are recognised and empanelled by MeitY. In India, micro data centres lack visibility and are businesses usually being operated by smaller enterprises while larger data centres are backed by well established conglomerates.

Government policies for DCs recognise facilities with more that 5 MW loads. Smaller outfits or micro data centres lack visibility and recognition even inside the government and are often excluded or generalised from data centre definitions.[18]

Micro data centres with capacity up to <1 MW could make for a viable business model and we are seeing micro data centres and edge solutions that can host 25-300 kW of capacity emerging in India. There is a need for these centres to be recognised and supported as a 'category of providers' for them to avail data centre policies and tax subsidies.

Overview of use cases that are suitable for MDCs

Micro data centres (MDCs) may be small in size, but they are capable of handling significant workloads, including those requiring substantial GPU power. For example, a 300kW MDC, which can accommodate nearly 300 NVIDIA H100 GPUs.

Traditionally, India's compute infrastructure has been dominated by CPU-based workloads. However, as the country embraces AI, the demand for GPU compute is surging. AI adoption is increasing across startups, enterprises, research institutions, and even government sectors, leading to an inevitable need for scalable GPU resources. While India's National Supercomputing Mission has made strides, it has underestimated the rapid growth of AI and its associated compute requirements. Here, MDCs offer a solution that is cost-effective, resilient, secure, and customisable to meet these evolving needs.[19,20]

MDCs can be tailored to balance both CPU and GPU workloads, adjusting configurations based on specific location and user requirements. This flexibility makes MDCs ideal for a range of applications, from everyday enterprise IT tasks to the more demanding needs of AI-driven operations.

Real-time video processing at edge

Real-time video processing for security cameras in smart cities, campuses, or retail environments at edge data centres can reduce the latency. Immediate threat detection and response without relying on distant cloud servers.

For example, BlazeWatch AI is an advanced fire detection system that leverages AI technology integrated with existing CCTV infrastructure to provide reliable and timely fire alerts. The system operates through a multi-stage detection and verification process, including preliminary colour detection, video classification, and contextual analysis using generative AI models. The use of Johnaic, an edge compute solution equipped with both

[18] response to draft data centre policy, 2020
[19] What are GPUs? Why is India scrambling for them? - The Economic Times
[20] National Super Computing Mission | Department Of Science & Technology

GPU and CPU capabilities, significantly enhances the system's real-time processing efficiency. Von Neumann AI's Johnaic accelerates the video classification and generative AI models, ensuring that the BlazeWatch AI system can quickly and accurately verify fire events, reduce false positives, and trigger appropriate alerts.

This integration of AI-driven capabilities within the BlazeWatch AI system exemplifies how micro data centres can support critical workloads, such as real-time video processing and safety monitoring. By utilising edge computing with Johnaic, the system minimises latency, ensuring rapid response times, which is crucial for fire detection and safety management in environments like schools and other sensitive facilities.

Healthcare: Balancing privacy and efficiency

Healthcare data transactions require a high level of trust between data centres, healthcare providers, and users. MDCs, with their privacy and resilience, offer an ideal solution. While some hospitals have shifted to cloud-based systems, many remain cautious due to concerns over PII (Personally Identifiable Information). MDCs provide a middle ground, offering the benefits of cloud storage while maintaining the security and control of on-premise solutions. For example, they can facilitate the secure storage and rapid retrieval of patient data, support AI models for predictive diagnostics, and ensure low-latency access to critical medical imaging data.

Radiology-AI

One of the prominent examples of AI applications in healthcare is computer aided diagnosis or CAD on radiology imaging in Tuberculosis. India has seen 28.2 lakh cases with a 12% mortality rate in 2022, the highest in the world today. AI interventions at many stages of TB have been deployed by companies like Qure.ai through their products.

Qure.ai is able to determine TB in patients from their chest scans. These solutions enable access to quality and rapid diagnosis at the last mile. Accessibility to such AI solutions could potentially be limited by the availability of compute infrastructure in more remote regions. The infectious nature of TB makes it important to treat it early and correctly. By using AI, the scans can be read and diagnosed much faster and secondary tests can be performed on the patients in a few hours to confirm a positive diagnosis. These models are now used as monitoring tools, detecting TB in patients who were screened for other concerns and improving the detection of TB in cases otherwise missed.[21,22]

The introduction of AI to brain imaging[23] has also been making medical treatments faster by up to 40% by providing results in 3min. These scans have made diagnosis of strokes more confident in early stages and allowed the teams around stroke emergencies to mobilise faster.[24] Tuberculosis and strokes are just one instance of how radiology has been moved by the introduction of AI. It might be misleading that an AI model of such calibre is expensive to

[21] https://qure-website-images.s3.ap-south-1.amazonaws.com/Medica_x_Qure_Whitepaper_Data_driven_insights_The_role_of_AI_in_radiology_s_evolution_192a69bfa1.pdf?updated_at=2024-06-26T05:57:31.853Z
[22] https://qure-website-images.s3.ap-south-1.amazonaws.com/TB_Diagonsis_South_Africa_6a664381d4.pdf?updated_at=2024-07-18T09:32:41.315Z
[23] https://us12.campaign-archive.com/?u=03cab87922cef7331fbb43d1e&id=b3ce2e6c09
[24] https://www.qure.ai/product/qer

use are requires premium infrastructure. This is not the case, as demonstrated by Qure.ai who are able run these models in remote regions of Africa on a single laptop and with limited power

Compute requirements for Inferencing

Qure.ai inferences a model- nearly 8B parameters or ~13GB in size on a single 12400F Intel chip. The model can inference entirely offline and only needs network connections to connect to the X-ray machine and to upload the scan to the cloud.

Edge solutions are optimal to low latency inference. Each scan is nearly 5MB and takes 1 min to run the model on. The Qure.ai solution has written up to 200 patient reports through AI, limited only by the number of patients in the camp.

Compute requirements for Training

To train its nearly 8B parameter model, Qure.ai took a batch training approach, starting with a cluster of ~40 Nvidia RTX 4090 GPU's for training- supported by their micro data centre in Goregaon, Mumbai.

Edge computing solutions like Johnaic, are able to provide low latency and offline responses in remote regions and make healthcare more accessible. Micro data centres serve as the perfect solution to bootstrap startups like Qure.ai in their early stages by providing on premise, easy to maintain and affordable compute.

Edge data centre for healthcare at India's edge: 10BedICU case study[25]

The 10BedICU initiative has established over 200 ICU units across 9 states in India, significantly improving critical care delivery in underserved areas. Currently, the initiative operates on a hub-and-spoke model, with tertiary hospitals serving as hubs and 10BedICU units in district and sub-district hospitals acting as spokes. The system relies on state-deployed cloud infrastructure (GCP, AWS, or Azure) to run the CARE (Comprehensive App for Rural e-health) software, requiring high-bandwidth internet connections (minimum 100 Mbps) for real-time video communication, patient data transmission, and remote monitoring.

Transitioning to edge Micro Data Centers (MDCs) could offer substantial benefits for the 10BedICU initiative. Edge MDCs, deployed at spokes or supporting multiple units in the vicinity, could reduce latency, improve reliability, and enhance resilience during internet outages. This shift has the potential to decrease long-term cloud expenses for states while enabling local processing and storage of critical patient data. Additionally, edge infrastructure could support the deployment of AI models like CARE Scribe and CARE Discharge Summary directly on edge servers, further enhancing healthcare delivery capabilities.

Implementing edge MDCs would require adjustments to the current infrastructure setup. While maintaining the hub-and-spoke model, each spoke would need local servers and LAN wiring to connect bedside equipment and high-resolution network cameras. The 10BedICU team would be responsible for installing and configuring these systems, including setting up

[25] https://ohc.network/

servers, testing local networks, and integrating devices with the CARE software. Staff at both hubs and spokes would require training to effectively utilise the new edge-based system.

The potential shift to edge MDCs aligns with the 10BedICU initiative's goal of leveraging technology to improve healthcare access and quality in underserved areas. By enhancing reliability, reducing costs, and enabling advanced AI-driven services, edge computing could significantly augment the initiative's impact on critical care delivery in India. As the project continues to evolve, the integration of edge MDCs represents a promising direction for future development and expansion of the 10BedICU model.

Everyday enterprise needs

Most enterprises engage in database management and telecommunications. Maintaining databases of people, projects, and communication influx are all major workloads in today's data centres. These IT workloads can easily be handled by MDCs as demonstrated by Vigyanlabs which today handles much of the instant messaging data around the country.

Storage solutions

Enterprises also require robust, scalable storage solutions for archiving data needed for compliance, legal, and historical purposes. MDCs provide the high capacity and cost-efficient storage needed, with compute resources focused on data management, integrity verification, and retrieval rather than intensive processing.

Supporting public sector’s compute needs

The public sector, particularly in rural and semi-rural areas, can benefit greatly from MDCs. RailTel, for example, is exploring the deployment of edge data centres in 102 cities to support its ICT initiatives. Similarly, initiatives like MeitY’s MeghRaj show how public and private partnerships can support e-Governance through distributed compute infrastructure. In cases like Digi Yatra, where the scale of operations is predictable, MDCs provide a reliable, low-latency solution that can support AI-powered airport management systems across India’s vast network of airports.

AI Startups

MDCs offer AI startups a strategic edge by enabling localised, real-time data processing with significantly reduced latency. This infrastructure not only cuts operational costs compared to traditional cloud solutions but also allows for precise customization to meet the specific demands of AI workloads. Startups can maintain data sovereignty and compliance by processing sensitive information on-site, avoiding the complexities of transferring large datasets to centralised cloud servers. Furthermore, the distributed nature of micro data centres enhances system reliability and scalability, positioning startups to scale efficiently while maintaining optimal performance across diverse applications.

Supporting Tier II and III cities: Expanding access to AI

As AI adoption spreads to Tier II and III cities, the need for localised compute resources grows. Many of these cities host incubators and research centres, making them ideal

locations for MDCs. Large players like Nxtra by Airtel and Sify are already establishing data centres in these regions, recognising their untapped potential. We have identified around 119 Indian cities with incubation and/or academic centres as potential MDC locations.

Uniqueness and potential in Tier II and III cities:

- Nearly 50% of all MSME's are located in Tier II or III cities. These cities have untapped potential for growing markets and the economy expands and supports regional developments.
- There is incentive from the state governments to grow the infrastructure of cities and maybe establish tech parks and businesses in these regions.[26]
- They provide a unique opportunity to build green DCs with easier land acquisition to align with the industry's growing sustainability goals.
- Cities like Jaipur are already picking up in the AI jobs race.[27] Tailoring content to reach every city[28] and enabling AI in healthcare for the Tier II and Tier III cities are already seeing demand.[29] MDCs with AI compute will benefit from the AI incubators and research hubs residing in these cities.

BFSI: Enhancing security and speed

The BFSI sector, driven by initiatives like DigiDhan, faces an increasing need for efficient data processing. With the rise of digital transactions and growing data consumption, MDCs offer a viable solution by processing data at the edge, closer to the point of origin. This reduces latency, enhances the speed of operations such as fraud detection and real-time trading, and improves data security and compliance. Additionally, MDCs provide the scalability needed to handle traffic surges at a lower cost compared to traditional cloud services.

Today, a single server can handle north of half a million transaction requests a day. As chips get more powerful, less space is needed for the same transactions. Large conglomerates like JPMC back this with theirs understanding of transaction speed requirements. With more financial aggregators entering the space, a MDC is the perfect edge solution for a digital India's BFSI needs.

E-commerce

E-commerce today is decentralised by ONDC but is unable to keep up with low latency infrastructure needs. The narrative of a centralised data centre system serving a decentralised commerce hub can benefit from a decentralised compute network of micro data centres.

IoT, 5G, 6G and future use cases

5G is an artefact of the present, but the cutting edge network speeds are bottlenecked by server's performing data transactions at 100s of milliseconds as opposed to data speeds of

[26] Role of MSMEs in enabling growth in Tier 2 and Tier 3 Cities in India
[27] Artificial Intelligence: AI jobs rise in Tier 2 cities, namely - Jaipur, Indore, and Agra, followed by others, says GetWork report - Jobs and Career News | The Financial Express
[28] AI in healthcare will be a game changer in Tier 2 & 3 cities, Health News, ET HealthWorld
[29] AI in healthcare will be a game changer in Tier 2 & 3 cities, Health News, ET HealthWorld

1 millisecond per MB. Edge data centres are the solution to application latencies and with 6G around the corner MDCs are the future.
Autonomous vehicles have already taken off outside of India- and electric scooters in India have started to build edge cloud solutions like Ather stack. The advent of more software driving technology means low latency, real time and edge MDCs.

The future of compute is edge and India can lean into MDCs to decentralise and energise its compute- both existing and new.

Economics of a micro data centre

Traditional approaches to designing, building, and operating Data Centres have been associated with high capital expenditure (CAPEX) and operational expenditure (OPEX). Large data centres, often consisting of thousands of racks and consuming hundreds of megawatts (MW) of power, have been constructed to achieve economies of scale and improved efficiency. The Advent of AI workloads have had adverse impact on the environment as these Mega DC were never designed with sustainability.

However, the concept of micro data centres (MDC) has not been widely adopted due to the historically prohibitive costs. This document presents the cost structures and design elements of a Green Data Centre (Green DC), demonstrating how it is economically viable to construct a MDC. The same design principles can be applied to build larger Data Centres in a modular way, significantly reducing both CapEx and OpEx.

Design and sizing guidelines for a 20-Rack data centre

1. Total Rack Space: Designed to accommodate 20 racks.
2. Office Space: Includes space to accommodate up to 20 personnel.
3. Rack Population: Out of the 20 available racks, 8 are populated as follows:
    - 1 Communication Rack
    - 1 Patch Panel Rack
    - 1 Network Rack
    - 2 Compute Racks: Housing 24 servers.
    - 1 GPU Rack: Housing 8 GPU servers, each with 8 GPUs for a total of 64 GPUs
    - 2 Storage + Backup Racks
4. Network Operations Centre (NOC) and Security Operations Centre (SOC): Fully equipped setups included.
5. Multiple Internet Service Providers (ISPs): Ensures redundancy and high availability.
6. High Availability (HA) Design: All components are designed for HA, with no single point of failure.
7. Exclusion of Land Costs: Land costs are not included in the costing analysis.

By following these guidelines and principles, the Green Micro Data Centre not only becomes a cost-effective solution but also aligns with the sustainability goals necessary for modern infrastructure.

Investment- Capex and Opex model of a green data centre

The following table provides a detailed comparison of the Total Operating Expenditure (OpEx) and Capital Expenditure (CapEx) over a 5-year period for a Green Data Centre (Green DC) versus a Traditional Data Centre (Traditional DC) solution. All costs are presented in Indian Rupees (INR) in crores.

Description	**Green DC**	**Traditional DC Solution**
IT (Hardware, Software) *	₹ 13.65	₹ 58.87
Non-IT Equipment	₹ 7.92	₹ 14.28
Architecture, Design, and Project Management (PMC)	₹ 1.00	₹ 2.50
Bandwidth Cost	₹ 19.20	₹ 19.20
Construction of Building & Maintenance Charges	₹ 3.38	₹ 3.70
Fuel for Diesel Generator Set	₹ 0.01	₹ 1.49
Electricity Consumption	₹ 0.33	₹ 14.32
Manpower	₹ 14.54	₹ 27.10
TOTAL	₹ 60.03	₹ 141.46

***IT costs include the chip and software licence costs.**

- The GPUs selected are a combination of A2, A10, A40 and A100 models.
- Using open source software has made the green data centres' IT costs lower than traditional data centres. For a complete management stack for the CNS including hypervisor, operating systems, cloud stack for managing compute, storage, network, GPU, NOC - Network Management and monitoring tools, SOC - Security Operations Centre, other tools like SIEM, Vulnerability Analysis, PAM/PIM, AIM, threat hunting, Vulnerability Management, Patch Management, Cloud Orchestrator, a self- service portal for users to provision cloud infrastructure and backup and recovery tools have high costs if one purchases commercial licences.
- Traditional DC Cloud Design uses expensive infrastructure like fibre channel, SAN etc. which are eliminated in the green data centre solution. Instead green DC operators use a software-defined DC which is mostly servers and a few network devices only (router, firewall, switches). Everything else is software.

CapEx vs. OpEx

Category	**Green DC**	**Traditional DC Solution**
CapEx (One Time)	₹ 18.60	₹ 42.27
OpEx (Over 5 Years)	₹ 41.43	₹ 99.20

Environmental Impact

Metric	**Green DC**	**Traditional DC Solution**

CO_2 Emissions (Tons)	0	15,599
Equivalent Trees to be Sequestered	N/A	257,914 Trees (Grown over 10 Years)

Power Usage Effectiveness (PUE)

PUE	Green DC	Traditional DC Solution
Power Usage Effectiveness (PUE)	1.2	1.5

Additional Costs to Government/Utility

Cost to the Government/Utility for Setting Up Additional Power Plant or Grid Capacity and Maintenance	Green DC	Traditional DC Solution
Cost	Nil	₹ 6.00 Cr

Return on Investment

Green and AI-optimised micro data centres are nearly 1/3rd of the capital investment as a traditional data centre and have ½ the operational cost too. This allows for compute offerings to be priced much lower than the traditional CSP offerings and a wide range of pricing options can still create returns.

The above model is for 20 racks but the same can be scaled to 2 or 3 times and create even larger profits as the human resource input does not grow linearly with the data centre size. The use of AI for automation and monitoring of the data centre operations further reduces the human resource costs and creates higher returns on investment as the data centre scales up.

Financing- Capital and Investors

MDCs are seen as high risk investments, making it harder for them to acquire capital. The interest rates on the debt is high; recognition and subsidy by government intervention would make capital investments feasible to MDCs.

MDCs are the edge enabler of Mega compute and not competition

They are low risk in scale, flexibility on investments and sunken costs. MDC markets are not threatened by the market for mega data centres.
There is a market for MDCs as the last mile delivery of compute that are hyperlocal and distributed. Much like the energy market- there is space for massive solar farms and rooftop solar to both co-exist.

The lack of a standard MDC model hinders financing and capital acquisition

Banks find it difficult to finance MDCs due the the lack of a model, this is similar to challenges in lending to agriculture or SMEs. The default is a high rate of interest to save their investments without MDCs value on cash flows or discounting future cash flows. Calculating the value of an MDC is different from that of a mega data centre. While mega data centres optimise to be economies of scale, micro data centres optimise to be edge, renewable and scalable.

Initiative needs to be taken to recognise MDCs which can enable them to be seen as having great value in investments that they are as opposed to uncharted and risky business.

What does it take to build a MDC

These key decisions will have to be taken into account to construct, manage and operate a data centre.

1. Decision 1: Where could the MDCs be located?
2. Decision 2: What are the key requirements to build a Tier III DC in India?
3. Decision 3: Should one build with partners or alone?
4. Decision 4: What should the execution strategy be?

Decision 1: Where could the MDCs be located?

- Certain states (for eg, Tamil Nadu) are excellent candidates for building and running data centre(s) for a variety of reasons: ample network connectivity, proactive government policies/tax subsidies and power availability (off the grid and renewable energy sources) being the main ones.
- As suggested in the section on RailTel, attaching the expansion of micro data centres to growing government use cases like the Indian Railways could be a fast and effective way to grow to many Tier II and Tier III cities.
- Over a 100 cities in India have technology institutions and incubators driving the AI startups as seen here. Another case to push micro data centres for being located at Tier II and III cities.

Following global leaders in technology like China who have similar strategies- linked here. The location of data centres can be moved away from the populous urban areas and expand into the regions where development is yet to reach. These regions may also have the added advantage of accessible and cheaper renewable energy sources.

Decision 2: What are the key requirements to build a Tier III DC in India?

This section evaluates the infrastructure and location, approvals and hardware requirements for a GPU-centric data centre facility in India.

- Infrastructure needs like UPS, high capacity connectivity, land and water requirements, cooling support, disaster recovery and redundancy and Uptime security and compliance are needed.
- Skilled human resource and management systems are required. The location of MDCs in Tier II and Tier III cities extrapolates to more employment opportunities and better curriculums in these cities.
    - Running a Cloud Service operation requires adequate manpower with the requisite skills. However, with AI & automation a significant amount of Cloud DC operations can be centralised and can be managed remotely. Take the example of the Telcos – the towers (with radios, base stations) are automated with minimal local staff.
    - Currently India faces an acute storage of skilled talent to architect, design, build and operate Data Centres/Cloud Services.
    - The skills required for operating and managing a data centre are acquired on the job and no formal institutional degree or diplomas exist presently. To skill the incoming workforce and create the talent pool required to operate data centres, encouraging institutions to run degrees and diplomas in data centre management and operation skills is needed. Modifying the syllabus to include:
        - Data Centre Electricians, HVAC Technicians and Engineers, Network Technician/NOC Operations, SOC Operators, Security Engineers, AIOPS Engineers, Cloud Storage Engineers, DCIM Engineers, Data Centre Design Engineers and Architects, Cloud Architects
- Statutory Approvals from nearly 30 central and state government bodies are needed which tends to restrict the infrastructure investments to experienced groups like Yotta (by Hiranandani).
- Some of the key approvals and licences required include: Other Service Provider (OSP) licence, Cloud Service Provider (CSP) licence, approvals under various labour laws like the Factories Act and Shops and Establishments Act and approvals for land, power supply, water, and other infrastructure.
- Hardware is controlled by chip manufacturers like NVIDIA who don't seem to sell MDCs directly. OEM channels like Dell and HPE are then used to procure the hardware. Other chip manufacturers like AMD and Intel also offer hardware options.
- Software (GPU) licences are also needed which can be as expensive $22,500 and are valid for five years. Some data centres have considered building their own software stack but reliability and effectiveness of an open-source software stack will have to be further investigated.

Decision 3: Does one build with partners or alone?

The Indian compute provider market is complex and fragmented. Success requires experienced partners familiar with the local landscape. Key partner contributions include:

- **Technical Expertise**: Specialised knowledge in data centre infrastructure. Example: Vigyanlabs partnered with 6+ vendors to build a facility in 9-12 months.
- **Regulatory Compliance**: Help navigate approvals and requirements. Example: Yotta's collaboration with Hiranandani Group.
- **Logistical Advantages**: Local knowledge for site selection and infrastructure setup.
- **Operational Efficiencies**: Expertise in optimization, monitoring, and automation.

Decision 4: What is the execution strategy and estimated project duration?

The estimated build-out and setup time for a high-calibre compute facility with a minimum of 20 racks (accommodating 200-240 GPUs) is approximately 18-24 months.

- Land procurement, obtaining necessary approvals, building construction, and establishing electricity and network infrastructure are expected to take 12-18 months.
- The procurement and installation of IT equipment and software set-up are included within the total build-out time duration of 12-18 months.

Status quo of policies to support MDCs in India

Governing bodies like STPI, MeitY, NITI and even states have policies supporting data centre infrastructure and management. These policies were written keping larger data centres as the target and hence gaps must be filled to accommodate MDCs too.

Extrapolate STPI infrastructure to serve MDC infrastructure requirements

STPI has established 5 major data centre projects and provides Project Management Consultancy Services for government aided or private projects including the management of some data centres. STPI can provide its general infrastructure and high speed data communication services to the micro data centre ecosystem.

MeitY recognises the gaps in compute infrastructure, data, AI financing and institutional capacity

To address these issues the expert group has suggested establishment of IndiaAI centres with these objectives:

- To establish multiple centres of excellence that tackle critical societal problems and enhance foundational research
- Develop industry-academia startup ecosystem to develop technology
- Strengthen incubation facilities powering AI startups

IndiaAI Future Labs has been suggested to improve the infrastructure and compute capacity. Micro data centres can host the AI innovation hubs with Secure Distributed Data Grids across the country to support startups and build academic and industrial collaborations.

NIDHI schemes are incentives to bring compute to Tier II and Tier III cities

The establishment of Technology Business Incubators(TBI) that convert innovations to startups and Centre of Excellences(CoEs) to facilitate startups to go global have been incentivised by funding. Many of these TBIs, CoEs and STEP incubators are in Tier II and

Tier III cities, as seen in the spreadsheet here. The AI first startups of today will benefit from AI compute available at an MDC.

NITI Ayog has suggested building a Common Compute Platform and enhanced support for AI based startups. In its recommendation is the notion of building a *100 AI petabyte* infrastructure that is multi-tenant, energy saving, low latency and has a deep learning software stack called AIRAWAT- AI Research, Analytics and knoWledge Assimilation plaTform. This infrastructure alongside established incubation hubs and venture funding will support the advancement of AI based developments in speech recognition, NLP and more.

State policies are geared for DCs to move to non-Tier I cities and even incentivise Greener DCs[30]

State policies on IT and DCs from Maharashtra, Haryana, Karnataka, Odisha, Tamil Nadu, West Bengal and Uttar Pradesh have taken key policy initiatives such as:

- Stamp duty exemption
- Development/ FSI related Incentives
- Capital subsidy
- Electricity exemption
- Power subsidy
- Infrastructure support(including UPS and water supply etc.)
- Tax Benefits
- Green Incentives
- Ease of approvals

Massive investments in Maharashtra and Tamil Nadu and leading hyperscalers and cloud service expanding to emerging untapped markets such as Kochi, Jaipur, Ahmedabad, Lucknow, Patna and Visakhapatnam are indicators of the shift of data centre establishments to Tier II or III cities.

Recommendations: Our proposal for an MDC-forward data centre policy

Recognition of micro data centres as a “category of providers”

- We encourage MeitY to recognise and empanel MDCs as a category of reliable providers in the compute and data centre market. This recognition will help drive demand generation, policies, subsidies and investments towards MDCs
- Currently, the Indian government has granted infrastructure status to data centres with a minimum capacity of 5 MW IT load. To promote MDCs, lower the capacity to 25 kW to accommodate a special category of infrastructure - MDCs.

[30] https://mktgdocs.cbre.com/2299/b51f6965-a387-4e71-8f8d-b779d1ad734d-109929665.pdf

Simplify certification and compliance requirements for MDCs

A simplified indigenous version of DC Certification - perhaps by STQC would reduce a chunk of the capex and opex involved in getting these certifications and renewing them timely. To be affordable to MDCs, ideally it should cost less than 5 Lakhs per year.

Aid demand generation and investments towards MDCs

- Ease financing and capital investments through tax breaks like STPI regime for software exports, simplifying risk assessment & valuation for funding, new instruments bonds REITs or even building a "hybrid" funding ecosystem[cite]
- Currently govt mandates that all departments to procure 10% goods from MSMEs, they same thing can be extended to MDCs. 15 to 20 percent of government data centre/cloud needs can be routed to MDCs that have been empanelled and certified to meet security, reliability and availability guarantees of such workloads. Consider that requirements like Aadhaar can entirely be processed by one MDC as it fits to 70 racks within 1 MW. Supporting the growing MDC landscape would boost their profits as Udaan promoted smaller Tier II and III city airports.
- Eventually, 50% of the government's expenditure in data centres should go to MDC over the next 10 years.
- GST on all MDC related equipment be reduced to 12% since it is an essential service for Digital India.
- NIIF, LIC, PFC, Power Grid, SIDBI, other PSU banks, and similar sovereign funds to support MDC. Lending to be done at minimum collateral.

Driving Innovation and Make in India for a sovereign compute ecosystem

- Foster innovation by supporting the startups at its forefront. Allow for tenders to recognise the startups and diversify the constraints to allow for innovation in hardware, software and operation.
- Promoting more Made in India Software Stack and incentivising local Cloud/data centre Software Vendors can lower costs of the software (SaaS) layer.
- Accelerate the fix with the Production-Linked Incentive (PLI) Scheme 2.0 for IT hardware for enhancing India's manufacturing capabilities. An average incentive of around 5% on net incremental sales (over base year) of goods manufactured in India and covered under the target segment, to eligible companies, for a period of six years. This includes incentives for servers and chips designed and built in India.[31] Hardware PLI scheme is presently not accessible to new companies and cannot participate as per the rule.[32] It takes 1-2 years to set up a hardware manufacturing plant before making revenue and the above scheme prevents new entrants which needs to be addressed.

[31] Production Linked Incentive Scheme - PLI 2.0 for IT Hardware | Ministry of Electronics and Information Technology, Government of India

[32] "For application under Domestic Companies: Consolidated Global Manufacturing Revenue of the applicant (including its Group Companies) should be greater than INR 10 crore for the target segment or greater than INR 20 crore for electronics hardware products/ sub-assemblies/ components." Production Linked Incentive Scheme - PLI 2.0 for IT Hardware | Ministry of Electronics and Information Technology, Government of India

MDCs can accommodate AI compute

- Guidelines on PII, data security and privacy for the healthcare and BFSI sectors to use Gen AI services supported by cloud computing. Policies on localisation of citizen's data for high risk AI use cases.
- Policies to prevent AI vendor lock-in and encourage a diverse ecosystem of providers
- Incentivise green-powered AI and environmentally sustainable practices for training and inferencing in data centres
- LLMs, ML Models and GPUs need to be offered as 'services' and standardised by MeitY to ensure Government Departments can access Gen AI and AI Services to transform governance cost effectively.
- Standardisation of APIs used to access the compute through cloud and procedures for empanelment can ease application development, migration, security and resilience of data centres.

Fostering an inclusive ecosystem

- Recognise stakeholders in the MDC supply chain, champion customer personas, and consortiums like OCC.
- Including curriculum to create the operation, management and engineering expertise required to build and operate MDC. More details suggested here.
