# Mid-Q1CY24 update for [OCC]

Prepared by swaroop rajagopalan Tanuj Bhojwani Tanvi Lall project is hosted on github



## Vision

Open Cloud Compute" (OCC) is an open network for compute infrastructure to help meet the growing demand for compute, helping unlock access to compute at scale through a Digital Public Infrastructure (DPI) Approach.

**Why OCC?**

1. Address the growing compute demand; AI in the hands of a billion+ people requires significant computing power
2. Keep the markets competitive; majority market share of a few providers in the market limits access to cost-effective cloud compute, consequences will be slower progress on AI-led solutions being scaled for use cases in health, education, agriculture and climate.
3. Create a public good, democratising access to AI, compute infra across users, builders, infra-providers

**What is the OCC network?**

1. A discoverability interface for users of IaaS, PaaS, SaaS computing service providers
2. A set of interoperable APIs like the Beckn protocol; the project focuses on identifying and executing the foundational blocks required for such a network to be in place for commercial or non-commercial transactions

3. Build trust via certifications as verifiable credentials, registries
4. Recognises micro data centres as a category of providers and assists with financial, technical, infrastructural policy and incentives
5. Allows users, businesses to easily plug and play to use, based on their specific compute requirements

**How will the OCC network flourish?**

1. Democratise access: it'll be an open network built on standards, allowing easy access and discoverability for users and providers alike. This empowers users with more choice and affordability, and also spawns new independent compute infra providers, expanding capacity and fostering innovation. By leveraging open-source software, we'll encourage wider participation and reduce entry barriers
2. Financial incentives: we envision collaborations, partnerships across public, private organisations to incentivise, drive investments encouraging more players in building compute infrastructure, and help trickle down awareness, incentives to the end users for adoption of this infra to meet their computing needs

**What are we trying to do in the short-term?**

1. Focus on business use case that's +1
2. Desirability and feasibility intersection mapping for arriving at early use
3. Bring hero stories out to make it a 'smart' thing to do

## Project goals and history

**2024 Vision: Take the OCC network live and complete >1 transaction between a domestic provider and customer.**

**Q2 and Q3 | Jan - Jun 2024**

We are working towards two objectives in the H1-2024:

1. Establish a strong business case for an OCC network:
    a. Understand pain points that existing data centre providers and software service providers of this stack face.
    b. Identify use cases and applications that could be early adopters of this network.
    c. Identify incentive schemes that lower the barrier for providers and customers to join this network.
2. Design a technical architecture that gives providers and customers a positive NPS experience.

**Q1 | Oct - Dec 2023**

1. Concept paper circulated and feedback sought from compute infrastructure providers and customers in India.
2. An article highlighting this effort is published in Economic Times.
3. Concept was presented in a workshop at the Global Technology Summit in December 2023 in New Delhi.

## Open questions

In order to achieve this vision, we are trying to answer these questions:

1. challenges, pain points
    a. What are the challenges and pain points in the current paradigm from a provider and customer perspective? Decent understanding from survey and 1:1 conversations
    b. Costs of compute for training, using foundational models are high. What are the ways we can bring these costs down? Examples
        i. access to cheaper infra, discovery of price, infra, match supply to workloads
        ii. will better quality data reduce costs?
        iii. will specialised re-training or retuning of existing models help? Or smaller models that are niche/ focused at lower costs?
        iv. can having common shared best practices between training to inferencing (eg. open metadata git example) or using federated learning reduce costs?] Partial understanding
2. use cases, applications, user requirements Partial understanding
    a. If the OCC network was to be a reality, what use cases/applications would it service? *+1 thinking is being applied here on what incremental value add we can add & build on them to users or providers*; We have a list of ideas
    b. What are the low hanging fruits/quick wins that would help us validate the market for such a network? We have a list of ideas
    c. What metrics are important to customers that the network needs to absolutely solve for? For e.g., performance, reliability, security Partial understanding
    d. Performance vs reliability? Any best practices for optimising trade-offs? Partial understanding
3. open source, security Partial understanding
    a. Where does open source software help?
    b. Can open cloud compute infra be reliable, secure? Will users trust this infra pipeline with their data, code, proprietary information?
    c. Are there any security, privacy concerns in using open cloud compute platform? Especially for AI workloads?
        i. does technologies like federated learning, multi party compute address privacy?
4. interoperability, deployment considerations Partial understanding
    a. Any existing initiatives or frameworks in place ensuring interoperability and compatibility between AI infrastructures deployed? (like skypilot)
    b. Any considerations when choosing providers for open cloud AI infrastructure deployment?
    c. SAAS, PAAS, IAAS — big cloud providers have an ecosystem of applications built by themselves, and by others on their platform, how do we extend these use cases on open cloud?

## Understanding demand

We segmented customers using the three following categories:

1. Based on workload (traditional compute needs like databases, storage vs. AI experiments/workloads). *The developer community in India is also an interesting demand segment who want to run AI experiments but need access to infrastructure.*
2. Based on industry type (companies could be building technology products, larger enterprises in essential industries)
3. Based on where they are in their lifecycle (early stage, growth-stage startups, mature and established firms, mature and growing firms).

## Key takeaways *[Open questions in italics]*

1. Apart from costs and issues like data sovereignty, large global CC provide not only Compute, Network and Storage but also productivity applications that customers like and are familiar with. *If traditional compute in large CC providers work ok today. What can OCC do here?*
2. Early-stage companies prioritise agility, cost-efficiency (with lower headcount). PaaS and SaaS solutions offer pre-built infrastructure and tools, helps them to launch, iterate quickly without hefty upfront investments in hardware, software, and expertise, allowing them to focus on business. There are providers who also supply/offer bare metals (with a basic software service on top of that e.g. Vigyan Labs, Hydrahost) as an IaaS to customers. *From the pie of customers, how many prefer IaaS and what are they using it for and will having a Craigslist or map of all available domestic compute help?*
3. Discoverability is a challenge and needs to be solved for. Once the customer finds the provider, incentive schemes (like compute credits, third party audit/checks will help bring customers online).
4. Cost is a key driver from a customer perspective. But in the world of compute credits, they don't think about it till it starts hitting them. *In early days, should OCC focus on new (where priorities are not cost or growth-stage customers where the sovereignty/price/vendor lock-in pain point is understood but it would require migration/openness to being multi-cloud for OCC to flourish?*
5. For early stage customers focused on R&D and PMF, OCC would have to simplify the value proposition and likely develop specific SaaS offerings that would utilise domestic infrastructure and be easily deployable. *Are there 1-2 offerings that many companies are trying to experiment with and can we start with that?*
6. Trust in providers is key. *How can the network enable that?*
7. AI is a use case but is still new. *For AI, should the goal be to enable experimentation on OCC to enable larger populations of people to think about it?*

***We have focused our time on figuring out the +1 step that could go live on OCC from a customer and provider perspective.***

# Low hanging fruit that the OCC network could solve for

## In the world of AI

1. **+1 for domestic providers of infrastructure and specific SaaS/PaaS offerings:**
    a. Craigslist for available capacity (gpulist, GPU.net) that allows for discoverability and price transparency for bare metal access/ IaaS
    b. Aggregation of demand for open-source LLMs: If smaller players knew of sustained capacity that customers have for open-source LLMs either for training and/or inference, they could allocate a given amount of capacity to these activities.
2. **+1 for customers, AI on an API:** Enabling SaaS offerings on Indian cloud providers that helps give access to domestic infrastructure for specific use-cases relevant to India in the form of read.mes, APIs that can go live on OCC
    a. Examples: Creating product catalogs for e-commerce startups (Step-by-Step Guide Using E2E Cloud), Powering chatbots (step-by-step guide by E2E Network)
    b. Talk to some of the smaller providers so we can jointly host some READMEs and some setup guides, essentially, on how to run these LLMs on these providers. What that hopefully will do is
        i. Hopefully help with some discoverability
        ii. Help us understand where to focus on in the next phase, which is the more kind of the standardisation phase.

## In the world of traditional compute applications

1. Providing compute for storage on OCC
2. Enabling eSamudaay's hyperlocal marketplaces that need edge compute via OCC. There is a design workshop planned for mid March with Vigyan Labs and the OCC team.

## Survey results highlighting status quo when it comes to using compute infrastructure

*Cloud adoption survey* (10 minutes, focused on the organisation's compute practices, pain points, future requirements for compute infrastructure); *AI workloads survey* (5 minutes, focused on AI workloads and imagining a use case that could go on OCC from the organisation's point-of-view). Surveys completed by leaders at 35 unique organisations (HQ in India and US)

Report can be found here

### Insights from 1:1 interviews

At technology companies, pre-seed to seed stage, going for venture capital:

- These are typically early adopters of new technologies
- Leadership team is focused on achieving technology development milestones and product-market fit; don't want to be worried about infrastructure selection or management
- Avail compute credits in the range of USD 300-400K from global cloud service providers incentive scheme which can last them a few years (<3 years)
- Rely on engineering team to select infrastructure + services stack that they are comfortable with; don't want to rock-the-boat and introduce something new that the team has to then learn and is unfamiliar with
- Rely on SaaS offerings and not infrastructure directly. They don't want the hassle of deploying directly on servers, some of them utilise companies like Replicate to run their workloads in SaaS/PaaS formats.
- For example, in GCP, where customers can just give all the details and get a full cluster for training and they have already built all those pipelines to enable these services. But Indian cloud providers, there are no such pipelines available so customers have to get the clusters from them and manually manage every part of the pipeline service. This requires 3-4 engineering resources to manage this.

At technology companies, growth stage, raised venture capital:

- Compute costs become a concern as the business starts to scale.
- Egress costs hit hard. <We have heard domestic providers of cloud computing don't charge egress and that is why a few customer have migrated to them>
- Cost of compute infra is a primary driver in selecting the infrastructure vendor but it's not the only one, reliability, data security, network connectivity, single point of failure are also important factors.
- We heard from a few founders that they would be open to exploring offerings of domestic providers but just didn't know who was credible and would match performance of larger, global companies.

## Q1 Operational Updates

### Formed a team

A working group of 9 volunteers (5 actively contributing, 4 joining as listeners/sounding board) who meet once a week to review project content, discuss and decide on next steps in alignment with quarterly goals. Participants are senior professionals who have spent time in infrastructure-related roles at JP Morgan Chase, Akamai, NPCI, Amuse Labs, Amazon, clAppit.io, University of Maryland. We have two individuals who work in tech policy.

Advisors include Arvind Tiwary, Srinivas Varadarajan, Narendra Sen, team at Oracle Cloud Infrastructure, Mallari Kulkarni.

WhatsApp group for the OCC team: join by clicking here

## Talked to diverse customers

Our goal was to build a business case for OCC and identify specific use cases and applications of compute that could be serviced on it. In pursuit of this, we segmented the market for compute into two categories:

- organisations that actively build/run/explore AI applications and,
- essential industries' large enterprises who would like to store large amounts of data, access it frequently, differentiate their product offerings using data and have a wide customer base across India.

We then gauged customer interest via surveys and in-depth interviews.

1. Two surveys: *Cloud adoption survey* (10 minutes, focused on the organisation's compute practices, pain points, future requirements for compute infrastructure); *AI workloads survey* (5 minutes, focused on AI workloads and imagining a use case that could go on OCC from the organisation's point-of-view). Surveys completed by leaders at 35 unique organisations (HQ in India and US)
2. In-depth interviews (transcribed and insights here) with managers/decision makers of compute infrastructure: Spoken to *six* technology SaaS founders and *one* electronics manufacturing services enterprise firm.

## Customer landscape included in study (green = in-depth interview conducted)

| **For AI use cases and workloads** | | | | | |
|---|---|---|---|---|---|
| **Industry** | **Technology** | **BFSI** | **eCommerce** | **Healthcare** | **Agriculture and climate** |
| **GDP contribution** | | 5-7% (30% of workload is AI) | 4-6% | 4-5% | 15-18% |
| **Use cases** | AI applications - b2b and b2c | Risk assessment, fraud detection, underwriting algorithms | Consumer recommendations, customer support | Treatment and diagnostics, care management, administrative, productivity | Soil analysis |
| **Companies** | Atlassian, Zoho, e6data, Syde Labs, Neurobridge Tech | Bukuwarung, Mysa, Coverfox insurance, Jupiter | Swiggy, Flipkart, ONDC, Meesho, DMART | 5C Network, 10BedICU, Driefcase, Foster Health | Nurture farm |

| Non-AI use cases and workloads | | | | | |
|---|---|---|---|---|---|
| **Industry** | **Manufacturing** | **BFSI** | **CDNs and OTT** | **Healthcare** | **Trade and Transportation** |
| **GDP contribution** | | 5-7%, 70% of workload is non-AI | 1-2% | | |
| **Use cases** | ERP | Transactions, data processing | Content hosting, streaming, network management | Data storage | Inventory management, logistics |
| | | | Streaming | | Logistics tracking |
| | | | Network management | | |
| **Companies** | Dixon Technologies, Ati motors, Ather Energy, Atomberg | CTO @ Zolve | Akamai, aha, Viacom, Zee | Medibuddy, Practo, Apollo Hospitals | Jebu Ittiachen - Obvious Technology \| LinkedIn,, Geet Garg - Locus.sh \| LinkedIn |