MOCSI Compute Project: the compute build

Narrative to accompany the Compute Project deck. Draft, November 2026. Figures from model 0e13319.

Shared for feedback. This is not an offer of securities or a request for investment; please do not circulate it without MOCSI's permission.

Summary

MOCSI (Maine Open Compute Services Initiative) is building AI compute that public-interest institutions can trust: open-weight models we will host ourselves, on hardware we will run, in a data center where the host community is a stakeholder.

MOCSI has three parts: the Software and Services Project, the Compute Project, and the Facilities Project. The Compute Project is a liquid-cooled cluster of used NVIDIA datacenter GPUs, 64 at launch and growing to 96 by year 4. We seek **$2.14M in financing**: $1.42M for the launch hardware and $0.71M of working capital. Operations repay it over five years. Across six years the plan earns $11.55M in revenue and $2.15 for every dollar of operating and hardware cost (before loan interest, with the facility's capital funded separately), and it holds at least $1.16M in cash at every year-end after loan payments.

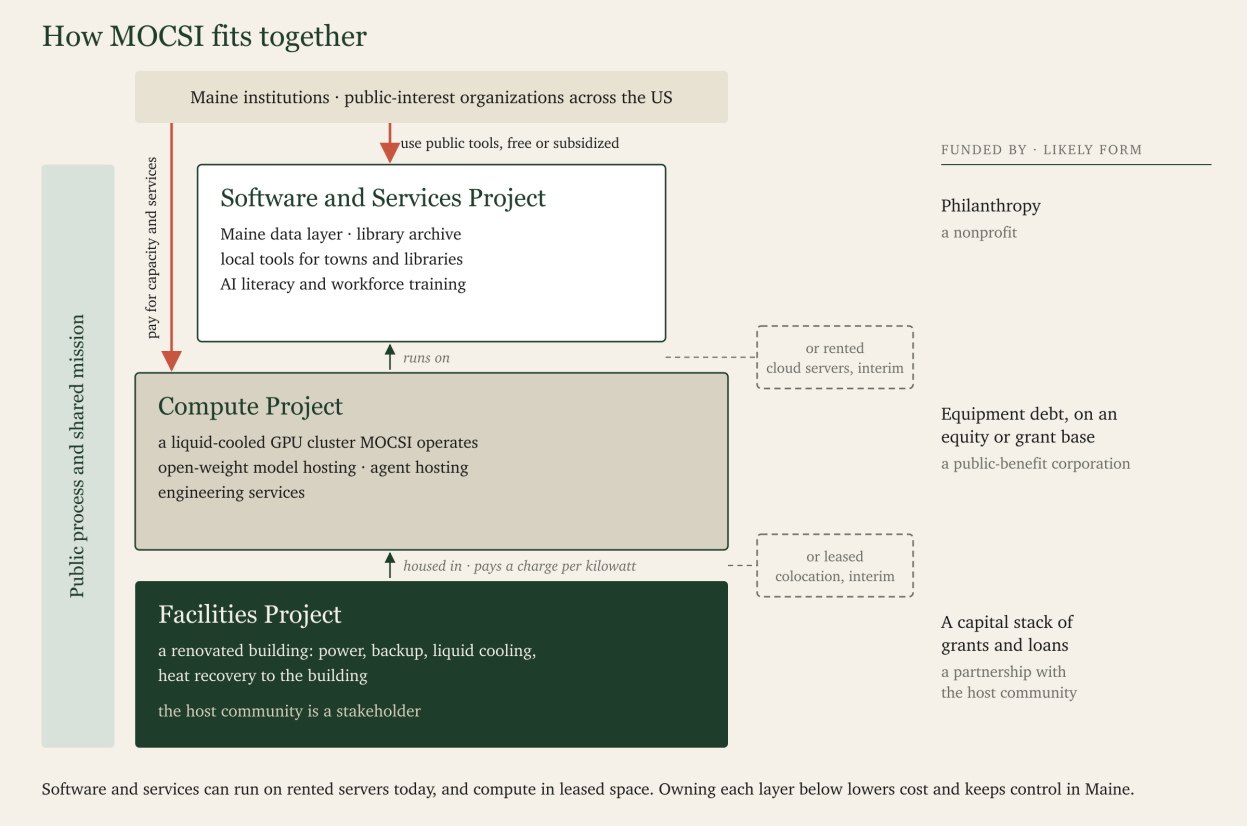


Figure 1: How the three tracks fit together. Each has its own funders and likely legal form, and each upper project can run on a rented substitute until the one beneath it is built.

The need

Community resistance cancelled or delayed $156B of data center deals in 2025, and Maine has seen municipal bans and a statewide moratorium bill. Meanwhile the institutions that serve Maine (towns, schools, libraries, and nonprofits) need AI they can inspect. The hyperscalers serve these buyers worst. A public university or

a town clerk needs to know which model answered, where the data sat, and who sets the price. Today they cannot get those answers from a provider in Maine.

What MOCSI is

MOCSI gives public-interest institutions affordable, auditable, non-extractive access to AI compute. Four commitments define it:

- **Open.** We self-host open-weight models, pin versions, and disclose model provenance.
- **Transparent.** We tell each customer exactly which model and version answered a request, where it ran, and who operates it. We disclose our jurisdiction and use no foreign subprocessors.
- **Auditable.** In a specific, narrow sense: the weights are published, we host them ourselves, and customers can confirm precisely which model ran. Customers also get contractual data handling and a right of exit with their data intact. Training data and training code are outside that promise, since very few competitive models publish them.
- **Community governance.** The community that hosts our data center will be a stakeholder in it, and a multi-stakeholder council will hold pricing and access decisions. A public engagement process with the host community will define the form of that stake.

Our pricing is competitive. Customers choose us for an inspectable stack, contractual data handling, and a seat in governance.

Who we serve

Two segments come first:

1. **Maine institutions:** local and state government, education (particularly higher education), libraries, and nonprofits.
2. **US public-interest nonprofits,** particularly those working on responsible AI, including members of the Pro-Human AI Coalition.

Maine is our home and first market. Our mission requires that a significant share of use come from Maine, starting at 10%. Excess capacity can go to for-profit customers, and public-interest customers always have priority.

We qualify the market honestly. Organizations outside the US are less likely to buy from us for data-residency reasons. AI safety organizations mostly need frontier models we cannot self-host, so they are unlikely to adopt early. Most Maine institutions lack the staff to run AI infrastructure themselves, so we pair every product with configuration and integration help.

What we sell

One GPU fleet serves four lines of business:

- **Reserved GPU capacity** (the fleet, less the GPUs held for seats). Dedicated NVIDIA GPUs on annual contracts for universities, research consortia, and organizations with their own engineering staff. Marketplaces sell spare hours.
- **AI assistant seats** (2 GPUs at launch, 10 by year 5, sized to planned seats). Hosted open-weight chat and document assistants for staff, sold as a subscription per seat per month. Institutions budget for subscriptions, and our market research found that seats earn more per GPU-hour than selling tokens.
- **Agent hosting.** Isolated, audit-logged sandboxes with a curated and supported set of open-source agent tools, running on CPU servers. This layer turns a single model call into a permit processor or a records clerk.
- **Engineering services.** Retrieval over an institution's own records (a records search, a permit assistant, a constituent chatbot), built as fixed-price projects with support retainers afterward. We deliver these with partners and contractors.

Nearby customers that run their own hardware can also rent rack space and power. The financial plan counts no revenue from rack space.

We do not train foundation models or host proprietary frontier models.

The compute build

The cluster runs on used NVIDIA datacenter GPUs in eight-GPU servers: H100s at launch, more H100s and the first H200s in year 2, and a second H200 server in year 4. Liquid plates on the chips cool them.

Buying used is a deliberate choice (D-20260928-06). Used datacenter GPUs cost a fraction of new ones, and the plan accepts three costs in return: less remaining life, a larger reserve for failures, and a lender who may shorten the loan to match the hardware's remaining life, or refuse used hardware as collateral.

	Launch (year 0)	Year 2	Year 4
Servers added	8	3	1
GPUs in service	64	88	96
Peak load	74 kW	101 kW	110 kW

Operating cash flow after repayment pays for the later additions ($0.61M), without drawing on the working capital. Liquid cooling captures about 78% of server heat, and the host building can reuse it for heat once the water class is settled (at the class modelled, the loop runs too cool for its radiators, and the facility estimate carries a heat pump to bridge it). The compute is portable and can be financed as equipment. It depreciates over five to seven years.

The building upgrades form a separate sub-project, estimated at $2.05M mid (range $1.03M to $3.85M). The estimate covers electrical service ($695k), backup power ($443k), and a liquid-cooling plant with heat recovery ($910k). The upgrades improve the host's building and stay with it, so we will fund them with the host community by grant, loan, or a mix. The compute pays the facility a charge per kilowatt of design capacity from the first year: $598 a year if a grant covers the upgrades, and $2,541 if a 20-year loan at 8.5% does, a rate that assumes a loan guarantee is available.

The design carries two sources of power (the utility and a generator), six dual-corded racks, a modular UPS with one spare module, and two physically diverse fiber entrances onto a ring, so a power module, a rack feed, or a fiber cut can happen without stopping the service. The transformer, the switchboard, the transfer switch, and the generator are each single. We priced a fully redundant electrical path and judged it not worth its cost at this scale, and we would rather say so than imply an availability we have not bought.

Financial plan

Revenue grows from $1.47M at launch to $2.37M in year 5. Services (seats, agent hosting, and engineering) grow from 6% of revenue at launch to 61% by year 5.

$ thousands	Y0	Y1	Y2	Y3	Y4	Y5
GPU capacity	1,387	1,161	1,422	1,166	1,115	919
AI assistant seats	55	166	331	552	773	966
Agent hosting	0	39	39	77	77	116
Engineering services	29	97	156	227	307	369
Total	**1,471**	**1,462**	**1,948**	**2,022**	**2,273**	**2,370**

Rental revenue eases over time as contract prices step down on aging hardware. Seats grow from 200 in year 0 to 3,500 in year 5.

Six-year revenue is $11.55M, and every dollar of operating and hardware cost returns $2.15 ($1.73 if the facility is financed), before financing costs. We tested fleets from 4 to 16 servers. Every size returns more than it costs, so the fleet can grow with demand. 12 servers fills the facility's 110 kW design.

The ask and repayment

Use	Amount
Launch hardware: 8 liquid-cooled NVIDIA H100 servers	$1.42M
Working capital: the launch ramp, before customers pay in full	$0.36M
Working capital: cushion	$0.35M
Total financed	**$2.14M**

At 8.5% over five years (a rate that assumes a loan guarantee is available), repayment runs $542k a year, with $575k of interest in total. The amount is the smallest loan that keeps cash at or above the $0.35M cushion at every month-end of the launch year. After payments, cash never falls below $1.16M at year-end and reaches $4.68M by year 5. If the facility is financed too, the amount rises to $2.26M and year-5 cash falls to $3.34M.

What would strengthen the plan

Each of these lowers the financing the compute build needs or the risk a lender carries:

- a guarantee of the equipment loan from an institution with a strong balance sheet, which the 8.5% rate already assumes;
- hardware donated by a nonprofit or by NVIDIA;
- philanthropy covering operating cost for the first year;
- a grant for the facility upgrades, which the plan already assumes.

Operations and staffing

A small core team runs the service. Volunteers, interns, and contractors cover software selection, maintenance, and application development, and a delivery partner handles larger engineering projects. The financial plan carries a staffing floor, maintenance and support hours, insurance, administration, tooling, and a reserve for failed hardware.

Next steps

1. Secure the site commitment with the host community.
2. Close financing for the compute build and a funding plan for the facility.
3. Install the hardware.
4. Install the software stack.
5. Onboard the first customers.

Status

These are our plans and principles. We are still developing the legal entity, the ownership structure, and the exact service configurations, and we will settle them with our host community, our customers, and our counsel. The financial figures come from a six-year planning model at mid-case estimates. Customer volumes (seats, projects, and agent-hosting servers) are estimates, and letters of intent will firm them up. Where something is undecided, we say so.