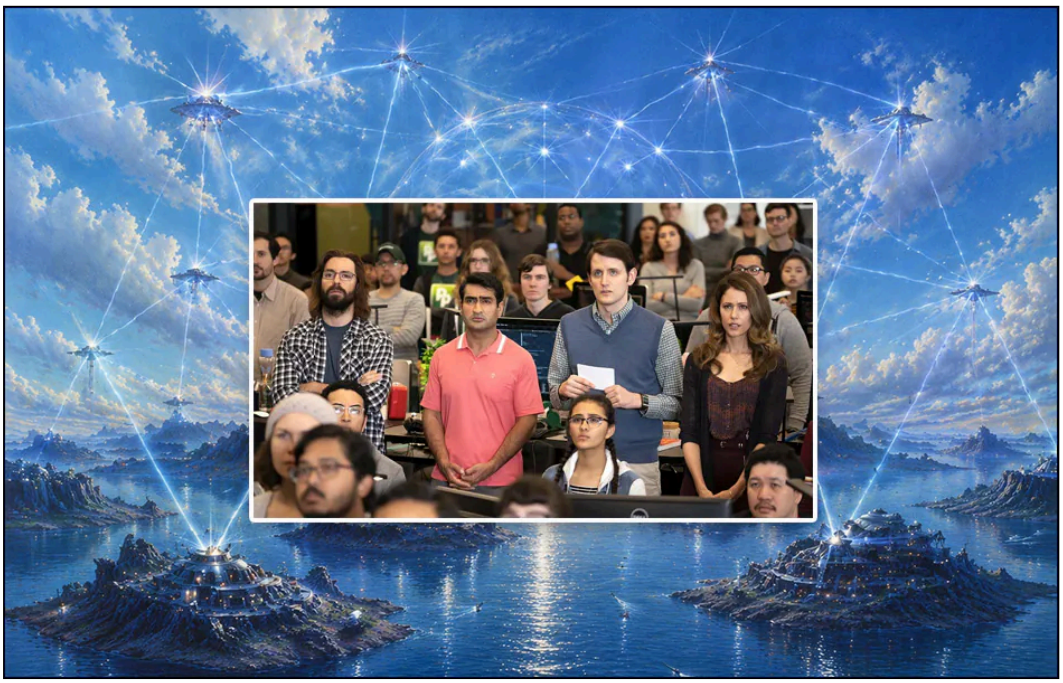
PiedPiper-Style Decentralized Inference Services for AI?

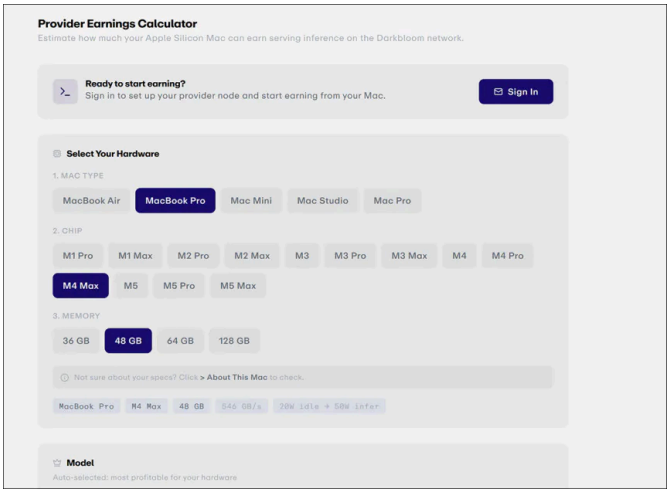
Originally published at https://www.unite.ai/piedpiper-style-decentralized-inference-services-for-ai/



Is 'BitTorrent for AI' an imminent possibility?

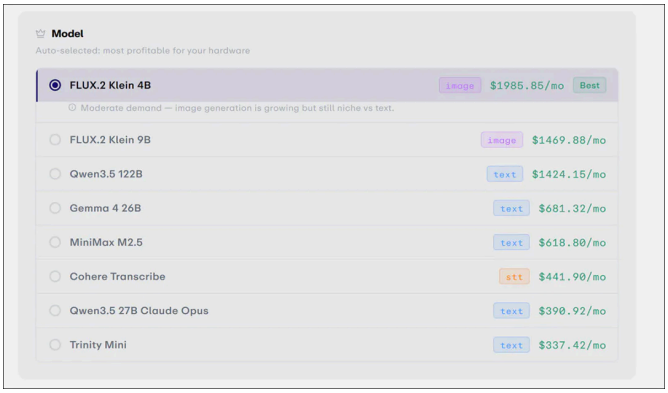
Opinion Having just yesterday finished a re-watch of Mike Judge's entertaining and acerbic tech-bro satire *Silicon Valley* – in which a group of socially-challenged geek geniuses attempt to create a 'new internet' called *PiedPiper*, via a mesh-network installed on everyone's mobile phones – I was interested to see the HN community engaging with a new offering of a similar nature.

Eigen Labs' DarkBloom sits somewhere between the egalitarian notion of a decentralized mesh network for AI inference, and crypto-mining profit motives, allowing the owners of Apple Silicon Mac systems to turn their equipment into an inference node:



From the Earnings section of the DarkBloom website, users can select which equipment they wish to rent out, and which AI models they wish to support. Source

The system concentrates currently on text-based models such as the agentic Trinity Mini (3B) and Cohere Transcribe, though it also offers diverse image-generating models such as FLUX 2 Klein 4B:



The range of models from which the 'landlord' can choose to rent out, together with monthly projected earnings.

Users participating in the scheme can apparently earn enough money in a solid month of inference provision to fairly regularly add a new Mac to an ever-growing chain, until, in theory, they can earn a full-fledged inference farm.

Effectively, a scheme of this kind that truly were to gain popularity (it has a cold start issue at the moment) might put enthused casual users back into a hardware-seeking stance, as in the last great cryptocurrency boom (and subsequent crash).

Not So Fast

However, for the little guys, that boat may have sailed. Besides AI's apocalyptic need for RAM, demand for global AI-enabled data center outfitting continues to raise hardware and services costs for the ordinary consumer, who previously had been able to monopolize RAM for crypto-mining, due to the peripheral nature of the activity, as well as regulatory uncertainty, which kept business interests circumspect on crypto.

While the super-cheap MacBook Neo has emerged as a crunch-beating alternative to ever-escalating hardware, its A18 mobile phone chip and 8GB of VRAM don't put it into serious contention as an inference machine.

But even if the end-user is not seeking to start a full-fledged inference farm, and merely wants to rent out their current unused M[n] capacity, the potential earnings seem significant, if the cold-start issue (an initial lack of users at the opening of a concern that relies on a high volume of participants) quickly resolves, and if the platform begins to advertise itself as something more than a curious experiment in potential demand.

Infer Different

Though a number of commenters have recognized a PiedPiper/Torrent-style democracy in DarkBloom's scheme, inference tasks are not as easily divisible as fragmenting a movie file into multiple hashed slices, so that it can later be reassembled in a torrent client.

The DarkBloom model is not proposing that a participant's M[*n*] chip handle *x*% of an inference task. In mainstream use, only a small number of frameworks or methodologies can achieve this kind of cross-GPU utilization on a single inference task, including NVIDIA's TensorRT LLM, which uses pipeline parallelism; and DeepSpeed's sharded inference which leverages model parallelism (MP).

Rather, your DarkBloom-enabled Mac would download and fire up one of the listed models and perform 100% of inference for paying users, with end-to-end encryption, and with prompts decrypted only on hardware-attested nodes, meaning that providers would not be able to read data during execution. The workload itself would constitute one or more text-based inferences, or at least one complete image.

It's not clear how extensive a single user session would be; as it stands, AI hobbyists are used to securing a GPU via inference farms such as RunPod; though it can take a while to secure the desired GPU at peak usage, the user gets to monopolize it as long as the session is not allowed to expire.

So it's possible that a single paying user could end up using a single rented DarkBloom Mac's M-series AI capabilities for a very long session, unless there is some logistical or compliance advantage in churning the clients between requests.

Macs have been singled out for this approach, apparently, because there are only a limited number of possible technical configurations for a participant, and it's therefore easy to assign apposite-sized models to a client.

Additionally, Macs capable of contributing to a DarkBloom network have a hardware secure enclave that guarantees a wall between the user and supplier.

These are all factors not so easy to rationalize across more generic, custom-made setups, and across the hundreds or thousands of known laptop/desktop Windows and Linux machines available over the last 6-7 years.

However, it must be obvious that the far larger non-Mac hardware pool could accommodate huge demand if their diverse characteristics could be rationalized, instead of – as with DarkBloom – hitching a ride on Apple's AAPL 0.36% limited spec-sets, which makes for an easy business proposition, and for a (presumably) far easier architectural approach.

Legal Oversight?

Perhaps the biggest issue facing a 'democratic' solution of this kind is the closed nature of the proposed process; governments around the world are currently engaged in new legislation that would effectively end internet anonymity wherever instituted, and are clearly not in a pro-privacy mindset in this period.

Therefore the prospect of random AI inference being carried out without filters, checks or balances, across a distributed network (if you can call DarkBloom that – it is more of an inference marketplace) seems, ironically, remote.

It's possible that DarkBloom, or other subsequent mesh inference schemes, will need to agree to backdoors that effectively restrict privacy to the *host*, who will not be able to see client jobs running; instead, the returned inference data would be made available through governmental agency man-in-the-middle (MiTM) structures, keeping all inference auditable.

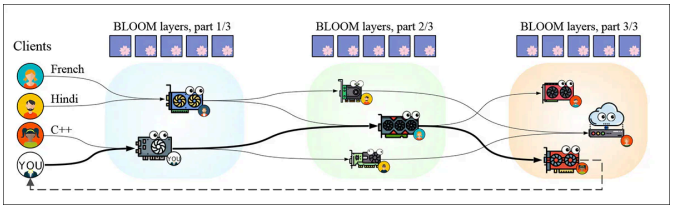
Presumably, if the rash of new laws proposing OS-level identity checks should ever achieve widespread adoption, such measures may become redundant. But without them, taking the current climate into consideration, a DarkBloom-style network would likely be considered akin to an AI 'darknet', where illegal AI-based activities could occur in secret.

Split Tests

To date there have been surprisingly few real attempts to do what a 'PiedPiper-style' system implies; in itself, DarkBloom sits at one extreme, distributing complete jobs to individual machines rather than attempting to fragment them across a network, while most production systems simply avoid the problem entirely by keeping inference on a single host.

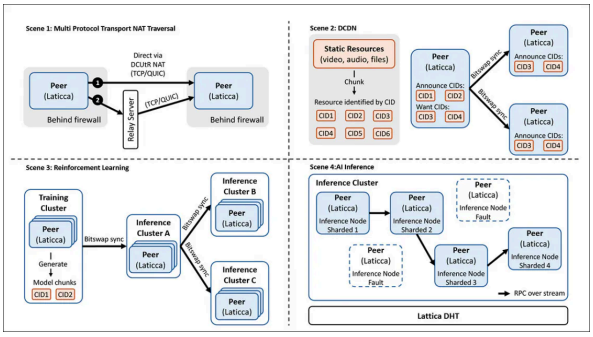
There are, however, a handful of projects that represent something a little closer to 'shared execution'.

Petals, which actively describes itself as a 'BitTorrent-style' network, distributes transformer blocks across multiple internet-connected nodes, passing intermediate states between them:



A typical Petals workflow, where a single inference request is routed across multiple remote GPUs, each holding a subset of model layers; unlike DarkBloom, execution is fragmented across the network, with intermediate states passed between independently-operated nodes, increasing latency and exposure at each hop while approximating a true mesh-style system. Source

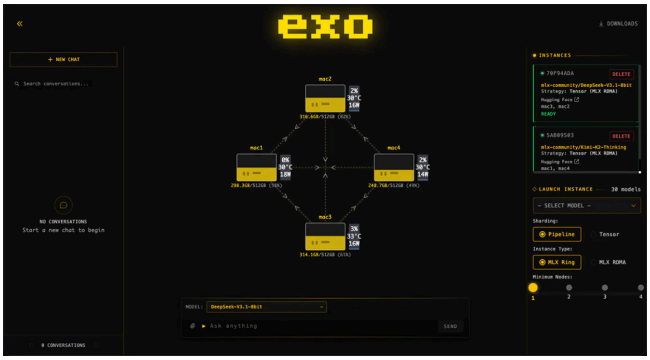
Hivemind experiments with similar peer-to-peer coordination and expert routing, though in the service of training models rather than inference from already-trained models; and Lattica focuses on the underlying networking layer needed to make such systems viable:



A schematic of Lattica, showing a lower-level peer-to-peer substrate that handles NAT traversal, content distribution, and DHT-based coordination (Distributed Hash Table), with sharded inference emerging only as one possible application layer; unlike DarkBloom or Petals, Lattica does not define an inference system itself, but provides the networking and synchronization primitives required to build one. Source –

All three of these models approach the mesh ideal, but at the cost of latency, instability and exposure.

Conversely, exo keeps inference within a local cluster, using fast interconnects to divide workloads across GPUs, without relying on the public internet. In practice, this kind of setup behaves less like a distributed mesh and more like a single extended machine, though there is clear possibility to extend this approach over a wider network:



A cluster view from exo, showing a small ring of local Apple Silicon machines jointly hosting a single model, with pipeline or tensor sharding distributing layers across nodes; unlike WAN-based systems, exo relies on fast local interconnects, effectively turning multiple devices into a single composite inference machine. Source

Finally, several commonly-cited approaches do not address inference at all: the now-venerable (2016) Google FedAvg; MIT's 2018 outing SplitNN; and the 2020 Australian offering SplitFed, are concerned with training distribution or privacy-preserving data exchange, rather than serving live inference requests.

Since training is a far more resource-intensive prospect than inference, any networks that prove to be able to distribute such a load effectively, across clusters or nodes, could have a disproportionate share of hobbyist and business interest later.

Conclusion

Because much of the technology in *Silicon Valley* was wild invention, we do not know whether PiedPiper was truly hash-driven (i.e., dividing and distributing data into chunks, torrent-style), or whether it 'settled' a task or even a session on any one node at any one time, which is what DarkBloom does.

However, the current scramble to provide training and inference hardware at the data center level indicates that the provision sector is either expecting to serve everyone, RunPod-style, or is gearing up to the most lucrative enterprise-level tier provision – a tempting prospect undermined by the general lack of moats in AI deployment.

If mesh-inference becomes a reality, it's reasonable to expect that among the earliest attempts to leverage it will be from the incumbents, such as OpenAI and Anthropic, who could either deploy dedicated systems within a massive existing app-install base, or else collaborate on open source systems which are easy to install (since companies of this size and reach have the money and motive to streamline difficult installations of this type).

As to whether a more democratic, user-driven mesh network could emerge, a true AI equivalent to BitTorrent – a number of factors are ranged against it.

Firstly, the current global drive against encryption and anonymity could remove or undermine many or all of the mechanisms that make systems such as BitTorrent anonymous, such as end-to-end encryption and VPNs. Once the 'generic' encrypted streams concealing such protocols are open to inspection, new layers of oversight and prohibition become possible, and this may undermine the appeal of a DarkBloom-style system.

Secondly, emerging or proposed regulations against 'abuse' of AI, or against anonymous operation of open source frameworks, mean that the cost of compliance – trivial at enterprise level – would likely take any smaller players off the market.

Finally – the power of a major sector player to Embrace, Extend and Extinguish (EEE, as Facebook and Twitter arguably did with more *ad hoc* internet communities), means that the current major players can operationalize and streamline the mesh model to their own advantage, in a market where end-users are almost totally intolerant to any friction in adoption.

First published Thursday, April 16, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More