Business Report

26 April 2024

parallex

Accessible Computing for All

Team 51
Andy Liu (andyl24@seas.upenn.edu, M&T)
Vikram Bala (vikbala@seas.upenn.edu)
Ethan Chee (echee9@seas.upenn.edu)
Anirudh Cowlagi (acowlagi@seas.upenn.edu)
Christian Sun (chsun@seas.upenn.edu)

TA: Heena Nagda
Faculty Mentor: Sebastian Angel

Executive Summary

An emergent cliché over the last two decades is that data, not oil, has become society's most valuable commodity[1]. Still, reductionist as it may seem, there is truth to the notion.

The average internet user generates about 2 megabytes of data per second[2] (roughly 1 high-resolution photo, every second). Society deploys these immense volumes of data towards an incredibly wide range of tasks, from content recommendation systems to delivering business analytics. More recently, the advent of large-scale powerful generative models like GPT-4 or its variants are trained using datasets on the scale of the *entire* public internet domain.[3]

Of course, much like oil, data in its unrefined and unprocessed state is of little use. Unfortunately, much like oil, the ability to efficiently process and extract value from modern data volumes is also still largely restricted to a select few players. Cloud computing services claim to restore parity by making computational resources available over the internet. However, with the services themselves being operated by those few with compute ownership, access remains prohibitive and expensive. As such, **Parallex proposes a collaborative solution to make computing truly accessible to all.**

Our key insight is straightforward: idle compute is plentiful. Indeed, at any point in time, about 63.5% of worldwide compute sits unused.[4] Thus, Parallex sets out to accomplish the following:

1. *Incentivize* the use of idle resources, whether it comes from a single laptop, or a university lab's unused server.
2. *Provide* these resources on a common marketplace, at a fraction of the cost of traditional cloud computing alternatives, while retaining performance.
3. *Manage* the network to ensure the security, safety, and efficiency of computational workloads.

Thus, Parallex allows *all* users around the world – whether they are individuals, universities, or start-ups – to easily, securely, and efficiently run distributed computing workloads by connecting them to a network of idle resources.

No longer confined to the privileged few, Parallex opens the doors to a global network of resources, allowing computation to become a shared utility, leveling the playing field for innovation and progress.

---

[1] The world's most valuable resource is no longer oil, but data
[2] Data Generation Statistics
[3] Common Crawl
[4] Idle Computational Resources

Parallex Overview

Parallex provides a service that increases global accessibility to distributed computing by leveraging idle computational resources around the world. By incentivizing *providers*, or the owners of idle machines, to run jobs on their computers, we can enable highly-parallelized, distributed computation at significant cost discounts to commercial alternatives on the market (cloud providers). In particular, Parallex offers both the middleware that manages job execution on heterogeneous provider hardware and the central scheduler that coordinates and manages job execution between providers.

Please see the *Parallex Walkthrough* video.

Stakeholders

Academia

Users in academia are often bottlenecked by funding and thus restricted in access to sufficient compute power available for their research needs. With either greater funding or cheaper compute power, the productivity of academic labs is likely to significantly increase. Academic users are likely concerned primarily with cost, and the ability for the compute service to efficiently and reliably perform the user's compute requests.

Industry

Industry professionals are facing similar challenges to academics in lack of specialized hardware for computationally-intensive tasks. Either the company does not have sufficient resources to purchase this computing power, or computing power available is rationed amongst a small number of users at the same company. This is true not only for start-ups, but also at industry leaders like Google[5]. Among industries that require significant compute power, pharmaceutical and technology companies are the leading players. Industry participants are likely most concerned about the security of their data, then cost and processing efficiency.

Providers

Providers want to be compensated for their energy consumption. They also want to ensure that Parallex does not affect their daily usage, and they want to make sure that their computers are secure and will not be adversely affected by their contributions to the Parallex network. Their primary concerns are thus compensation and security.

Value Proposition

Parallex provides value by 1) increasing the accessibility of distributed computing and 2) improving utilization of existing idle compute resources. For providers, Parallex better utilizes their idle resources and compensates them at rates better than their energy consumption costs, creating value. For Parallex users, Parallex provides computing power at significant cost reductions to industry norms, with little trade-offs. This allows them to run

[5] AI Developers Stymied by Server Shortage at AWS, Microsoft, Google — The Information

bigger jobs or more jobs and improve the output of their activities, which are often bottlenecked by resource availability, particularly in light of GPU shortage[6].

Industry Overview

The industry in which Parallex operates can be considered a subsection of the cloud computing industry. It might be best represented as an alternative to serverless computing offered by Amazon (AWS) as Lambda or Google (GCP) as Cloud Functions. We expect rivalry among existing competitors to be high, due to low differentiation between different competitors. We expect bargaining power of buyers to be high by the same reasoning. Bargaining power of suppliers is low due to the low number of large players, but may be adjusted to medium due to shortage of supply. The threat of new entrants is low due to a high barrier to entry, and the threat of substitutes is also low, with no real alternatives besides self-financing hardware purchases.

Market Research

We expect high CAGR growth rate in serverless computing, to the rate of 20.8% CAGR over the next 5 years[7]. We expect the industry to grow from $9.3 billion USD in 2022 to $28.9 billion in 2028. Furthermore, the recent explosion in generative AI is a significant external driver of demand for serverless computing. Training and deploying large-scale generative AI models requires a significant amount of computing power. For instance, Llama-2-13B, a 13 billion parameter large language model, required >350,000 A100 GPU hours for training alone.[8]

Estimates for the Cloud AI market, which is a closely related complementor of our section of serverless computing, see the market growing at a CAGR of 35.8% to $887 billion by 2032[9]. In aggregate, we see strong, steady growth in the market driven by external factors (primarily AI) that are unlikely to slow within the next decade.

Customer Segment

We segment Parallex customers into primarily three groups: academia, start-ups, and established companies. Academic and start-up customers are likely to be primarily concerned with the cost of Parallex compared to alternatives. A close but slightly less important concern to these two segments is the security of their data while using Parallex. Established companies are likely less concerned with cost, but more so the security and integrity of their data. To them, the risk of negative press and consequences of data breaches and leaks are not worth cost savings and increased opportunities in cheap computing power.

---

[6] The A.I. Industry's Desperate Hunt for GPUs Amid a Chip Shortage - The New York Times

[7] Serverless Computing Market Size 2023 with a CAGR of 20.8% : Latest Growth Rate, New Development, Market Segment, Sales & Revenue, Global Demand and Regional Outlook till Forecast Year 2030 Research Report

[8] Llama 2 Model Card

[9] Cloud AI Market Size to Grow USD 887 Billion by 2032 at a CAGR of 35.8% | Valuates Reports

To this end, we have selected *academic* customers as the most suitable go-to-market customer. They tend to have strong integrity of using the platform for anticipated uses, and can act as both provider and consumer to test and validate the platform. Their concern about personal reputation reduces the risk of antagonistic behaviors. Despite their high concern for low costs, we believe that we can provide computing power and competitive costs for their needs.

Please see **Exhibit 1** for a comparison of the customer segments for a selection of go-to-market customer.

Competitors, Substitutes, and Alternatives

Parallex's largest competitor is cloud providers - namely Amazon Web Services, Google Cloud Platform, and Microsoft Azure. Each of these platforms holds their own version of serverless computing, such as AWS Lambda, where users can purchase computing power. However, cloud providers are oftentimes expensive, and there is a large shortage of specialized ML hardware availability. These companies furthermore all have businesses which demand a huge amount of AI compute, where much of their availability will be reserved for internal usage only - further constricting publicly-available supply.

A further alternative to Parallex would be a privately-hosted datacenter, purchasing the newest publicly for-sale hardware such as Nvidia's A100 or H100 tensor chips. However, this is almost always intractably expensive - running upwards of $700,000 minimum per order of compute chips combined with a long waitlist[10]. Specialized knowledge would also be required for maintenance, further adding to the cost. The average users neither have the funding nor demand enough compute for the economics of maintaining their own datacenter to make sense.

Finally, there exists open source software that hosts peer-to-peer computing. These broadly fall under two categories - incentivized and non-incentivized, wherein the latter is significantly more popular than the former. The non-incentivized options, such as Petals [11]and Hivemind[12], suffer from a lack of adoption and users on the system (due to their non-incentive nature). Petals, furthermore, is restricted to finetuning large language models only. BOINC[13] is an academic project, but is similarly restricted to very particular computation use cases, with a complex technical and logistical process for adding new applications. Incentivized cryptocurrencies like GridCoin[14] are built on top of the highly restricted set of BOINC use cases.

---

[10] The San Francisco Compute Company
[11] Petals. Run LLMs at home, BitTorrent style
[12] Hivemind: Decentralized Deep Learning in PyTorch
[13] BOINC
[14] Gridcoin

Cost and Revenue Model

Cost

Parallex's costs are twofold -1) running central command nodes (job scheduler and managers) 2) compensating provider nodes. Our central command node costs will be similar to comparable cloud hosting uses. We anticipate the majority of our cost structure to be from provider node compensation - which is itself broken into two categories. 1) We provide compensation proportional to the computing power provided to the network. 2) We draw from a lottery pool at predetermined periods of time. Entries into the lottery pool are granted to providers proportional to computing power provided. We believe a combined reward system will be most effective in attracting and retaining potential providers.

Revenue

Our revenue model is more straightforward. Users pay a fee proportional to the amount of computing power used by the Parallex network for the job they want to run (e.g. # of CPUs, GPUs etc.). The majority of this fee will be used to compensate providers, while Parallex retains a pass-through portion for facilitating and managing the job execution.

In practice, we intend to facilitate all transactions by tracking jobs using Parallex Compute Units (PCUs) which differentially weight the various forms of computational resources that may be utilized, in addition to incorporating metrics such as reliability and network costs.

Feasibility Analysis

Exhibit 2 walks through preliminary pricing methodology and cost-revenue analysis for a sample job and compares it to existing cloud providers and energy costs to demonstrate revenue model feasibility.

We see that running jobs through Parallex is 4x cheaper than a traditional cloud provider ($0.091 vs $0.034) , for the same job. Although some jobs may take longer than cloud providers, Parallex still offers substantial performance benefits through Ray's distributed computation platform *and* makes many more computation tasks accessible to users by providing a large pool of resources.

Furthermore, we prove that provider payouts greatly exceed energy costs from providing their compute by a factor of four, at $0.004 per provider versus an energy cost of $0.001 per provider.

We finally show that Parallex can maintain a healthy margin using a 30% passthrough fee for managing and executing jobs, while maintaining the above benefits to users and providers.

Appendix

Exhibit 1. Go-to-Market Customer Selection Comparison

Go-to-Market Customer Selection				
	Security Requirements	Willingness To Pay	Antagonistic Behavior Probability	Ability to be Provider
Academic	Low	Low	Low	High
Start-up	Medium	Medium	High	Low
Established	High	High	Medium	High

Exhibit 2. Pricing Methodology and Cost-Revenue Analysis for Sample Job

Sample Job Specification & Required Resources

PyTorch Training (Deep Network Classification) - CPU Only

Number Processes (# Workers * # CPU Cores / Worker): 32

Baseline: Amazon Web Services

4 m5.2xLarge nodes (32 GiB RAM, 8 CPU Cores)

Job Runtime (4 workers, 8 CPUs / worker): ~200 seconds = ~0.06 hours

Cost: 4 * ($0.38 / hour) * (0.06 hours) = **$0.091**

Job Pricing

Parallex Provider nodes with 6 GiB RAM, 2 CPU Cores

Require 16 Provider Nodes, 1 Command Node ($0.015 / hour)[15]

Job Runtime Estimate: ~1000 seconds[16] = 0.3 hours

Revenue:

[Provider Nodes] 16 nodes * ($ 0.006 / node / hour) * (0.3 hours)

+ [Command Node] 1 node * ($ 0.015 / hour) * (0.3 hours) = **$0.034**

Costs:

[Payout to Provider Nodes] 16 nodes * ($ 0.004 / node / hour) * (0.3 hours)

+ [Command Node Cloud Instance] 1 node * ($ 0.015 / hour) * (0.3 hours) = **$0.024**

Gross Profit: $0.01 / job (29 % margin)

[15] Assume command Nodes are shared across jobs and clusters and are traditional cloud instances – assume ~5 jobs managed per Command Node, enabled by a m5.large AWS instance

[16] Assume 5x slower due to increased node count, hardware processor speeds, inter-cluster latency etc.

Energy Analysis

Uptime Cost / Provider: (10 W / core) * (1 kW / 1000 W) * (2 cores / provider) * (0.3 hours) * ($ 0.18 / kWh) = $0.0011 / Provider

Payout = $0.004 / Provider > $0.0011 / Provider = Uptime Cost / Provider

Initial Takeaways:

1. (Feasibility) Running jobs through Parallex can be much cheaper than running job through a traditional cloud provider ($0.091 vs $0.034)
    a. Note: Jobs may take longer than cloud providers, but Parallex offers substantial performance benefits through the use of Ray's distributed computation platform *and* makes many more computation tasks accessible to users by providing a large pool of resources
2. (Feasibility) Providers receive payouts that greatly exceed energy costs from providing idle compute to Parallex network ($0.004 payout / provider vs. $0.001 energy cost / provider)
3. Parallex can maintain healthy margins (+29 %) by imposing a passthrough fee for managing and executing jobs