Extremely Far Data Computing

Guy Wilks
gwilks@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Nathan Serafin
nserafin@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jennifer Brana
jbrana@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Nathan Beckmann
beckmann@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Brandon Lucia
blucia@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Abstract

Moving large swaths of data has never been more important. As compute latency becomes increasingly dominated by data movement times, research has focused on how to reduce these costs. These data movement costs occur on every level of the hardware stack, from moving data between cores to moving physical SSDs between racks in data centers. Yet one domain remains unexplored: how to move data between continents with high bandwidth and low latency. Undersea cables, while cost effective and low latency, limit bandwidth. Now more than ever, novel ways to transmit data across long distances are needed.

We take lessons from recent work on near data computing to devise a new solution, showing that we can transform our data movement problem into a rocketry problem. We explain how this solution has far higher bandwidth than existing solutions while providing security guarantees, ultimately proving that it isn't computer science we should be advancing, it's rocket science.

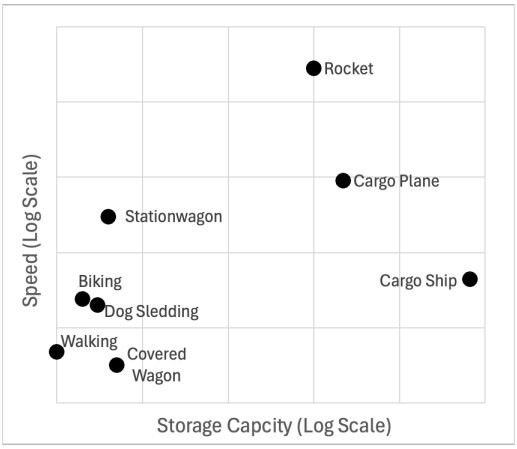

Figure 1. Storage Capacity versus Speed of Transportation for various methods of transporting compute

1 Introduction

The amount of data transferred globally (both data transfers between cities on the same continent, but also between different continents) is increasing at an alarming rate [8]. In 2018, Cisco predicted that global cloud data center traffic would reach 19.5 zettabytes by 2021 [2]. Due to the sudden increase in machine learning and artificial intelligence workloads, we have far outpaced this prediction [13].

The meteoric rise in data has created a problematic bottleneck in data movement [8]. Data movement refers to the shuttling of data at every level of the computing hierarchy, from cache lines being moved around processors to GPUs sending data over high-bandwidth NICs to data being transferred over long-range communication networks. As compute bandwidth and density continues to improve, more and more communication bandwidth is needed at every level of the computing hierarchy [7].

In response to this growing data movement problem, various solutions have been proposed at the chip architecture and datacenter scales. At the chip scale, architects have observed that Network on Chip (NoC) bandwidth can only scale so much, so instead of sending data across the chip, send the compute to the data, processing data either in memory [5, 9] or wherever the data is in the on-chip memory hierarchy [7, 12].

Solutions at the datacenter scale range from finding ways to route data more efficiently with smart NiCs [6] and specialization [1] to physically moving high density SSDs around the datacenter using vacuum tubes [8].

While these solutions solve local data movement problems, global communication, of which 95% moves via undersea cables [3, 10], remains an open issue. The MAREA cable, which connects mainland Europe to the United States, is 6,605 kilometers in length [14] and has a data transfer rate of over 26 terabits per second [11]. In total, about 300 cables spanning over 1 million kilometers provide three petabits of data transfer capacity per second [4]. As compute capacity increases, and as the amount of data being computed on continues to grow, these cables will fail to provide necessary bandwidth. Laying down hundreds or eventually even thousands more of these cables both impractical in nature. Additionally, these cables incur significant security risks, as severing the cables reduce inter-continental bandwidth

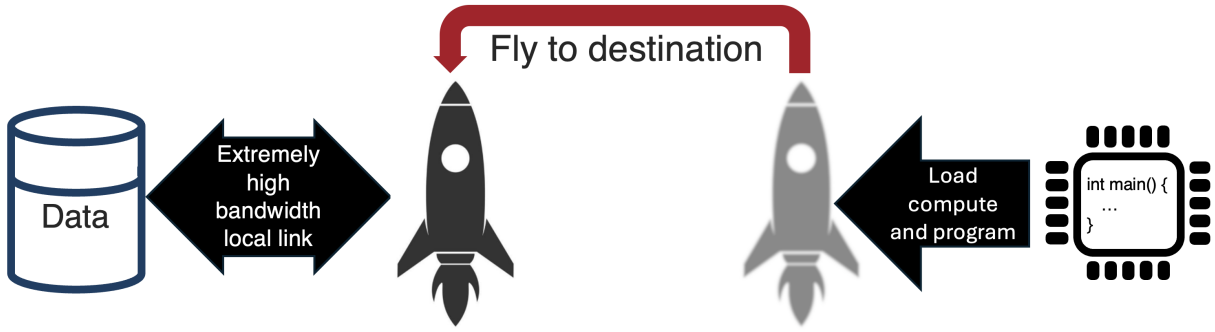

Figure 2. An overview of our proposed system

needed to support applications used by millions on a daily basis.

We can separate the types of data requirements needed by applications who utilize these undersea cables. In one camp we have applications that require low latency but also low bandwidth, on the other hand, there are applications that require high bandwidth but not low latency. Finally there are applications that require both or neither.

We take a lesson from near data computing in order to provide a solution for transferring data belonging to applications requiring high bandwidth but not low latency. **Instead of moving the data to the datacenter, we propose moving the datacenter to the data.**

2 Moving the Datacenter to the Data

By moving the data center to the data, the bandwidth of the system becomes limited by whatever the bandwidth of the local link it between the moved data center and the data, which will far exceed the allotted bandwidth from an undersea cable.

So the question becomes how to transport the compute. We look at two parameters for different transportation methods. The first parameter is latency, which is measured by the speed of transport. The second parameter is storage space, which is the measured volume of the various kinds of transport. Figure 1 shows the three operational points which we consider. Of course, each method of transporting data has its benefits and downsides. While using rockets are the fastest, cargo ships have the most carrying capacity. Using a stationwagon, bike, dogsled, covered wagon, or walking hardly help if the data is across an ocean. Additionally, planes, ships, and rockets cost significantly more than cars, bikes, or sleds.

We can also make quantitative bandwidth comparisons between different methods of transportation. The distance between Los Angeles and London is 8,756 km as the crow flies. It would take a single stationwagon (assuming no stops, and that it could go straight and over ocean) nearly 30 hours at max speed to travel the distance. A rocket could travel the same distance in less than an hour. Additionally, the carrying capacity of a station wagon is one 250th of the rocket. Meaning you would have to deploy 250 station wagons in parallel to provide the same compute bandwidth as one rocket. Because of this stark difference, we advocate for rockets, being the optimal balance of speed and capacity.

Figure 2 shows how this system might operate. To begin, load the compute, which could be any hardware configuration such as CPUs and GPUs, onto the rocket. Optionally you could either load the program on now or find one at the destination. Next, the rocket launches, flying to the destination at breakneck speeds. Once the compute has reached the destination, an extremely high bandwidth local link is set up between the source of the data and the compute on the rocket. Lastly, the rocket executes the available program on the data before launching off again.

3 Security

Rockets, in combination with the near data computing paradigm, present a unique security opportunity. Since the data never moves, it cannot be intercepted by someone monitoring network traffic, resulting in fully on-premises computing. That being said, we acknowledge that in reducing the data security risk we do introduce a national security vulnerability in sending rocket technology potentially all over the world. We expect and encourage all entities using this technology to consider that risk. Additionally, extra care should be taken to ensure that no programs loaded onto the rocket's computing system can affect the code responsible for scheduling the rocket.

4 Conclusion and Future Directions

As we have shown, resources dedicated to computer science research surrounding bandwidth intensive applications should be transferred to research for rockets which can carry more at faster speeds even cheaper. We believe this technology will eventually allow a more scalable and connected world than ever though possible. In conclusion, computer science literally is rocket science.

References

[1] Alvarez, C., He, Z., Alonso, G., and Singla, A. Specializing the network for scatter-gather workloads. In *Proceedings of the 11th ACM Symposium on Cloud Computing* (New York, NY, USA, 2020), SoCC '20, Association for Computing Machinery, p. 267–280.

[2] Cisco. Global cloud index projects cloud traffic to represent 95 percent of total data center traffic by 2021, 2018. Accessed: 2025-03-20.

[3] Federal Communications Commission. Circuit status report, 2025. Accessed: 2025-03-20.

[4] Flexential. Subsea cables: Key to a more connected world, 2023. Accessed: 2025-03-20.

[5] Kang, H., Gibbons, P. B., Blelloch, G. E., Dhulipala, L., Gu, Y., and McGuffey, C. The processing-in-memory model. In *Proceedings of the 33rd ACM Symposium on Parallelism in Algorithms and Architectures (SPAA '21)* (2021), pp. 1–12.

[6] Lin, W., Shan, Y., Kosta, R., Krishnamurthy, A., and Zhang, Y. Supernic: An fpga-based, cloud-oriented smartnic. In *Proceedings of the 2024 ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA '24)* (2024), pp. 1–12.

[7] Lockerman, E., Feldmann, A., Bakhshalipour, M., Stanescu, A., Gupta, S., Sanchez, D., and Beckmann, N. Livia: Data-centric computing throughout the memory hierarchy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems* (New York, NY, USA, 2020), ASPLOS '20, Association for Computing Machinery, p. 417–433.

[8] López-Paradís, G., Hair, I. M., Kannan, S., Rabbat, R., Murray, P., Lopes, A., Zahedi, R., Zuo, W., and Balkind, J. The case for data centre hyperloops. In *2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)* (2024), pp. 230–244.

[9] Mutlu, O., Ghose, S., Gómez-Luna, J., and Ausavarungnirun, R. A modern primer on processing in memory. In *Emerging Computing: From Devices to Systems – Looking Beyond Moore and Von Neumann*, K. Kim and R. Kumar, Eds. Springer, 2021, pp. 113–137.

[10] National Oceanic and Atmospheric Administration. Submarine cables, 2025. Accessed: 2025-03-20.

[11] Optical Fiber Communication Conference. Researchers break efficiency record for data transfer in ultra-fast transatlantic cable, 2019. Accessed: 2025-03-20.

[12] Schwedock, B. C., and Beckmann, N. Leviathan: A Unified System for General-Purpose Near-Data Computing . In *2024 57th IEEE/ACM International Symposium on Microarchitecture (MICRO)* (Los Alamitos, CA, USA, Nov. 2024), IEEE Computer Society, pp. 1278–1294.

[13] Statista. Volume of data created, captured, copied, and consumed worldwide from 2010 to 2028, 2025. Accessed: 2025-03-20.

[14] TeleGeography. Marea submarine cable, 2025. Accessed: 2025-03-20.