Natural Language Processing (2026), **32**, pp. 224–231
doi:10.1017/nlp.2026.10017

CAMBRIDGE UNIVERSITY PRESS

EMERGING TRENDS

Emerging trends: Supercomputers and clouds are so last century

Kenneth Church[†]

The Institute for Experiential AI, Northeastern University - Boston Campus, USA
Email: kenneth.ward.church@gmail.com

(Received 3 February 2026; revised 3 February 2026; accepted 6 February 2026; first published online 3 March 2026)

Abstract

Over the next decade, I expect more and more computing on phones and small computers and less on clouds and supercomputers. Supercomputers are super impressive to engineers, but not to economists because economies of scale have more to do with the size of the market than the size of the machine. Clouds are like wire-wrapped Cray computers; they were never designed for mass production. There was never much supply or demand. Supercomputers and clouds are too expensive and burn too much power. The future is more promising for phones because the market is larger. In addition, there are a number of other advantages to computing on phones and small (commodity) machines designed for mass markets: privacy, power, size, weight, latency, bandwidth, and especially affordability.

Keywords: Supercomputers; clouds; phones; economies of scale; commodity; power

1. Introduction

I have been working recently at VecML, which is deploying AI Agents, RAG, and chatbots everywhere, including small machines (phones/laptops) and big clouds.

Why compute on a phone (as opposed to a cloud)? I have lived long enough to experience many changes in the field, including several booms and busts (AI Winters and AI Summers), as well as swings back and forth between

1. computing locally (PCs, LISP Machines,[a] Thinking Machines[b], and Sun workstations)[c] and
2. computing remotely (clouds, time-sharing, and batch on mainframes[d] and slurm clusters).[e]

The current generation of students has grown up with few options other than cloud computing, but I expect that will change soon. Phones and small machines have a number of advantages over supercomputers and data centers: privacy, power, size, weight, latency, bandwidth, and especially

[†]I am also at VecML.com in Bellevue, WA, USA.
[a]https://en.wikipedia.org/wiki/Lisp_machine
[b]https://en.wikipedia.org/wiki/Thinking_Machines_Corporation
[c]https://en.wikipedia.org/wiki/Sun_Microsystems
[d]https://en.wikipedia.org/wiki/IBM_mainframe
[e]https://en.wikipedia.org/wiki/Slurm_Workload_Manager

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

affordability. Over the next decade, I expect more and more computing on phones and less on clouds.

There is a narrative that clouds will cook the planet, but I expect OpenAI will run out of capital before their emissions become comparable to cars. While data centers are not green, the world will move to alternatives such as phones, not because phones are greener, but because they are cheaper.

As for privacy, most LLMs are trained on public documents that can be crawled from the web, but many of the documents that users really care about are private. If you want to search photos/files on your phone, do you really want to upload your private data to who-knows-where?

As mentioned above, affordability is the primary concern. There have been suggestions that clouds are cheap. They are not. The parking fee to park a machine in a production data center for a month is about the same as the cost of the machine. I learned this rule of thumb in the 1990s at Bell Labs, but it continues to hold today. For example, I can rent a terabyte (TB) of disk from AWS for a month for about what it costs to buy the disk from BestBuy. Clouds may be convenient, but that convenience comes at a huge cost.

In (Church, Greenberg, and Hamilton 2008), we compared costs of data centers with alternatives based on commodity components (shipping containers and even condos). The last author, James Hamilton,[f] currently an executive at AWS, was an early advocate of replacing brick-and-mortar data centers with shipping containers. Our discussion of condos is a tongue-in-cheek suggestion intended to call attention to the advantages of general-purpose commodity solutions. Just as general-purpose micro-processors won out over special-purpose VLSI chips, so too, it is almost always cheaper to take advantage of economies of scale and commodity components designed for mass markets. Shipping containers are cheaper than brick-and-mortar, and phones are even cheaper than shipping containers.

2. The "Data center" in my home

Recently, I moved a website[g] from AWS to a "machine room" in my house. The "machine room" is shown in Figure 1. The rest of the "data center" in my home is shown in Figure 2.

The website is based on the "Better Together"[h] JSALT-2023 team and their github.[i] JSALT secured funding to host the website at AWS for a while, but when that funding ran out, I needed to find a cheaper alternative. As it turned out, the "data center" in my home was so much cheaper that the disk capacity could be expanded from 2 TBs to 68 TBs (for a one-time cost of $11 per TB).

The Mac Mini in Figure 2 has more GPU resources than I realized (and more than I could afford to rent from AWS). If more GPUs are needed, I can buy an affordable box from NVIDIA that is similar to a Mac Mini in size and cost (roughly comparable to a month's rent for an apartment in New York City). Since I care more about disk space than GPUs, I upgraded the Mac Mini with 4 TBs of SSD and 24 TBs of USB disk. Those upgrades cost about $500.

Working with VecML on phones, I am learning that phones also have more GPU-like capabilities than I realized. In addition, phones will improve more quickly than clouds (and small computers) because of economies of scale. Improvements in computing are often formulated in terms of Moore's Law. Moore's Law was originally motivated based on growth rates of transistors/chip, but these days, improvements in computing have more to do with economies of scale than transistors. The future is more promising for phones because phones have a larger market.

[f] https://uwaterloo.ca/math/profiles/james-hamilton-mmath-97
[g] http://recommendpapers.xyz/
[h] https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html
[i] https://github.com/kwchurch/JSALT_Better_Together

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press



Figure 1. The "machine room" in my house consists of a modem (left), plus a $500 NAS box with 4 disks (right). There is (consumer grade) electricity and network, but no chilled water, cooling, UPS, battery/generator backup power.



Figure 2. The "data center" in my house, a Mac Mini with a 4 TB SSD on top, and (too many) wires to monitors, cameras, microphones and speakers and 24 TB of USB disk (not shown).

3. Diseconomies of scale

I became interested in economies of scale based on conversations with Danny Hillis, a classmate and founder of Thinking Machines (Hillis 1985). At Thinking Machines, he thought he was building a supercomputer to compete with Cray[j], but he soon discovered he was losing to PCs. Smaller machines were not only cheaper than bigger machines but smaller machines were improving more quickly than bigger machines. He referred to this observation as a "Diseconomy of Scale."

When I took Danny's argument back to an economist at Bell Labs, I learned that economies of scale have nothing to do with the size of the machine and everything to do with the size of the market.

[j] https://en.wikipedia.org/wiki/Cray

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

There is a discussion of diseconomies of scale in Wikipedia.[k] Hamilton, the last author of (Church *et al.* 2008), wrote a blog post on diseconomies of scale 20 years ago.[l] He concluded that computing is cheaper at home than in the cloud. His post starts with a discussion of economies of scale. Clouds have advantages because large companies have considerable leverage when placing large orders. They may be able to negotiate better terms than consumers can expect at home.

> *The services world is one built upon economies of scale. For example, networking costs for small and medium-sized services can run nearly an order of magnitude more than large bandwidth consumers such as Google, Amazon, Microsoft and Yahoo pay. These economies of scale make it possible for services such as Amazon S3 to pass on some of the economies of scale they get on networking. . . These economies of scale enjoyed by large service providers extend beyond networking to server purchases, power costs, networking equipment, etc.*

But Hamilton quickly pivots from economies of scale to diseconomies of scale:

> *Ironically, even with these large economies of scale, it's cheaper to compute at home than in the cloud. . . Data centers are about the furthest thing from commodity parts and I have been arguing that we should be moving to modular data centers for years. . . Modular data centers help but they still require central power, mechanical systems, and networking systems and these systems remain expensive, non-commodity components. How to move the entire datacenter to commodity components? Ken Church. . . makes a radical suggestion: rather than design and develop massive data centers with 15 year lives, let's incrementally purchase condominiums (just-in-time) and place a small number of systems in each. Radical to be sure but condo's are a commodity and, if this mechanism really was cheaper, it would be a wake-up call to all of us to start looking much more closely at current industry-wide costs and what's driving them. That's our point here.*

The blog continues with a cost comparison between condos and Microsoft's data center in Quincy Washington.[m] Hamilton and I had visited that data center soon after it opened. The visit made it clear that power was the dominant cost, but not for the reasons that most people talk about. We are more concerned about the lack of commodity parts, upfront capital, and reliability requirements:

1. lack of commodity parts: custom-built brick-and-mortar data centers are more expensive than commodity components designed to serve multiple needs for larger mass markets
2. large upfront capital investments scale with peak capacity, as opposed to
   a. just-in-time investments that can be spread over time
   b. monthly expense bills that scale with consumption
3. unreasonable reliability requirements: It is very expensive to backup the power grid with batteries and generators. Many/most use cases do not require 99.999% uptime (defects per million).

Hamilton posted two more blogs on this subject.[n]

[k] https://en.wikipedia.org/wiki/Diseconomies_of_scale

[l] https://perspectives.mvdirona.com/2008/04/diseconomies-of-scale/

[m] https://en.wikipedia.org/wiki/Columbia_Data_Center

[n] https://perspectives.mvdirona.com/2008/09/embarrasingly-distributed-cloud-services/; https://perspectives.mvdirona.com/2008/10/hotnets-2008-paper/

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

4. Power

Much has been written about power and data centers and sustainability (Patterson *et al.* 2021, 2022; Ren, DiWang, and Sivasubramaniam 2012; Kong and Liu 2014; Lacoste *et al.* 2019; Lottick *et al.* 2019; Strubell, Ganesh, and McCallum 2019; Schwartz *et al.* 2020; Henderson *et al.* 2020; Sharir, Peleg, and Shoham 2020; Wu *et al.* 2022; Luccioni, Viguier, and Ligozat 2023; Zhao *et al.* 2022; Bouza, Bugeau, and Lannelongue 2023; Luccioni and Hernandez-Garcia 2023; Li, Jianyi Yang, and Ren 2023; Zhang, Wang, and Wang 2011; Stewart and Shen 2009; Liu *et al.* 2012; Li, Qouneh, and Li 2012; Goiri *et al.* 2013; Zhang, Wang, and Wang 2012; Deng *et al.* 2014). Much of this work assumes that the cost of power scales with consumption (as opposed to peak capacity).

Does it make sense to reduce consumption by putting the clouds to sleep when they are idle? Since the business case is constrained more by capital than expense, and more by peak capacity than consumption, once a firm invests in a data center, there are strong incentives to run the data center as close as possible to maximum capacity.

The business case for data centers is like the business case for airlines. If an airplane is about to take off, it makes more sense to sell the last seat at a discount than to remove the seat cushion to save weight. Putting the clouds to sleep is like removing the seat cushion. Once you have invested in a data center with the capacity to burn x watts, you have to run the data center as close as possible to x over the projected lifetime of the data center (typically 15 years). If there is a risk of stranding capacity, that is, running the data center at less than x, it makes more sense to sell the last watt at a discount before the plane takes off, or else the idle capacity will be wasted.

Much of the literature above focuses on training, a one-time cost, though inference will have more impact on carbon emissions, when AI applications become more widely deployed, because inference is a recurring cost.

Since I grew up in Rhode Island, which used to have lots of water mills like Slater Mill,[o] the idea of moving data centers closer to the power source has a romantic appeal. Hydroelectric dams are modern versions of water mills. Since it is so expensive to ship electricity from the dam in Quincy Washington to Seattle, why not move the demand for power closer to the source?

But when we visited the data center in Quincy, it became clear that while the romantic green story may have been great PR, it was not a great business case. While the power from the hydroelectric dam may be green, there is nothing green about batteries and generators, whether you use them or not. More seriously, the monthly power bill is a round-off error when compared to the massive upfront capital investment required to build the data center in Quincy. There may be plenty of skilled electricians in Seattle, but not in Quincy. It was necessary to recruit electricians with specialized skills from elsewhere, and build a place for them to live before work could start on the data center. Moreover, even in Seattle, there are relatively few electricians with the skills to work on a data center. It is relatively easy to find electricians that have worked on condos, but not so easy to find electricians that have worked on data centers.

The monthly bill for the power actually consumed was tiny compared to the massive capital investment. Why is the monthly power bill so small (and so irrelevant)? Suppose the power bill in Quincy is about 1/4 as much as in Seattle

1. because of shipping costs (perhaps 1/2 of the power will be lost in transit), and
2. the town's willingness to offer incentives to increase the tax base.

But even so, if one had to pay a mortgage for the capital investment, the mortgage would be larger than the monthly power bill. The business case just does not make sense, even with a great deal on the monthly power bill.

[o] https://en.wikipedia.org/wiki/Slater_Mill

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

As for reliability, it is common for a data center to overbuild the power grid with backup batteries and generators. As mentioned above, you hope you never need to use the backup batteries and generators, but they are expensive and dirty, whether you use them or not.

A few people backup the power in their home with batteries and generators, but most people do not. The supplier of the NAS box in Figure 1 recommends installing a UPS (uninterruptible power supply)[p] to give the NAS box enough time to shutdown gracefully in a power outage.

I do not own a UPS, though perhaps I should buy one since we have had a few outages this year. Since the UPS costs about as much as a disk, I have decided to use the cost of the UPS as "self-insurance." That is, I have decided that I would rather buy a new disk after a crash than to pay for a UPS now to avoid a future crash. Thus far, there have been no disk failures. I can always buy a UPS in the future if "self-insurance" turns out to be a bad investment.

When I worked for AT&T, we designed the network for 99.999% reliability (defects per million), but after deregulation, the market made it clear that customers were not willing to pay for that much reliability. In the case of my website, I would like the website to be more reliable than the power in my house and the network from my ISP, but affordability is a higher priority for me than reliability. Given the choice of running the website when there is power and network, or taking the website down (because I cannot afford to keep running it on AWS), I chose affordability over reliability. For many/most use cases, affordability is a higher priority than reliability.

For services like email, we suggested in (Church *et al.* 2008), replacing batteries and generators with geo-diversity. In this way, services can be designed to survive outages in one region by transferring load to another region.

While it may be romantic to move the data center closer to the power source, many data centers are located closer to the users because of the speed of light.[q] From this perspective, it makes sense to move the computation to the phone, which is even closer to the user.

5. Conclusion: Supercomputers (and clouds) are super expensive

My conclusion, that phones will beat out clouds, may be a hard sell, at least in the short term, especially to an audience of researchers that has grown up on cloud computing. Researchers tend to focus on training, where there may be some advantages to clouds, but going forward, if the business case for AI makes sense, then there will be more demand for inference than training. At scale, the cost of inference will dominate the cost of training since inference is a recurring cost that scales with usage, unlike training, which is a one-time cost. For many (simple) queries, much of the inference work can be done on the phone (at the edge of the network), and therefore, much of the load on the cloud can be moved to edge, where computing will be cheaper (and better in most respects), because of economies of scale.

Economies of scale have little to do with the size of the machine and everything to do with the size of the market. Supercomputers are super impressive to engineers, but not to economists. Supercomputers are super expensive and super heavy and burn too much power. Wire-wrapped Cray computers were never designed for mass production. There was never much supply or demand. In short, the business case for Cray computers was never super impressive, compared to other computer companies such as "the magnificent seven"[r] that have created a mass market for tech.

Clouds are sometimes referred to as supercomputers. Like supercomputers, clouds are super-expensive and burn too much power. On the other hand, the business case for clouds is better because the convenience makes it possible to support massive profit margins. My home "data

[p] https://en.wikipedia.org/wiki/Uninterruptible_power_supply
[q] https://perspectives.mvdirona.com/2012/07/why-there-are-datacenters-in-ny-hong-kong-and-tokyo/
[r] https://www.fidelity.com/learning-center/smart-money/magnificent-7-stocks

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

center" is cheaper than AWS because my home avoids AWS margins, as well as other costs associated with production data centers.

Eventually, the market will figure out how to build home data centers to provide the convenience of clouds at an affordable price. Reliability will likely suffer since reliability is expensive. For most use cases, the mass market will prioritize price and convenience above reliability.

I would like my website to be more reliable than the power in my home, but since I cannot afford what that costs, I prefer the unreliability of my home to the alternative (taking the website down).

I wish NAS companies did a better job with the box-opening experience. Turing on a NAS box is not nearly as much fun as opening a box from Apple. That has to change, or else clouds will continue to get away with embarrassingly large profit margins.

In retrospect, the market for phones has been more successful than the market for supercomputers (and clouds). About 10^2 Cray computers have been built, compared to 10^{10} smart phones. In aggregate, phones have more computational resources than supercomputers (and clouds). In short, since phones have a larger market, they have a more promising future.

References

Bouza L.D., **Bugeau A.** and **Lannelongue L.F.** (2023). How to estimate carbon footprint when training deep learning models? A guide and review. *Environmental Research Communications* **5**(11), 15014.

Church K., **Greenberg A.G.** and **Hamilton J.R.** (2008). On delivering embarrassingly distributed cloud services. In Hotnets, pp. 55–60.

Deng W., **Liu F.**, **Jin H.**, **Li B.** and **Li D.** (2014). Harnessing renewable energy in cloud datacenters: Opportunities and challenges. *IEEE Network* **28**(1), 48–55.

Goiri Í, **Katsak W.A.**, **Le K.**, **Nguyen T.D.** and **Bianchini R.** (2013). Parasol and greenswitch: Managing datacenters powered by renewable energy. *ACM SIGPLAN Notices* **48**(4), 51–64.

Henderson P., **Hu J.**, **Romoff J.**, **Brunskill E.**, **Jurafsky D.** and **Pineau J.** (2020). Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* **21**(248), 1–43.

Hillis W.D. (1985). *The Connection Machine*. MIT press.

Kong F. and **Liu X.** (2014). A survey on green-energy-aware power management for datacenters. *ACM Computing Surveys (CSUR)* **47**(2), 1–38. https://api.semanticscholar.org/CorpusID:12337968

Lacoste A., **Luccioni A.**, **Schmidt V.** and **Dandres T.** (2019). Quantifying the carbon emissions of machine learning. ArXiv abs/1910.09700.

Li C., **Qouneh A.** and **Li T.** (2012). iSwitch: Coordinating and optimizing renewable energy powered server clusters. *ACM SIGARCH Computer Architecture News* **40**(3), 512–523.

Li P., **Yang J.**, **Islam M.A.** and **Ren S.** (2023). Making AI less 'Thirsty'. *Communications of the ACM* **68** (7), 54–61.

Liu Z., **Chen Y.**, **Bash C.**, **Wierman A.**, **Gmach D.**, **Wang Z.**, **Marwah M.** and **Hyser C.** (2012). Renewable and cooling aware workload management for sustainable data centers. In *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE Joint International Conference on Measurement and Modeling of Computer Systems*, pp. 175–186.

Lottick K., **Susai S.**, **Friedler S.A.** and **Wilson J.P.** (2019). Energy usage reports: environmental awareness as part of algorithmic accountability. ArXiv abs/1911.08354.

Luccioni A. and **Hernandez-Garcia A.** (2023). Counting carbon: A survey of factors influencing the emissions of machine learning. ArXiv abs/2302.08476.

Luccioni A., **Viguier S.** and **Ligozat A.-L.** (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research* **24**(253), 1–15.

Patterson D., **Gonzalez J.**, **Holzle U.**, **Le Q.**, **Liang C.**, **Munguia L.-M.**, **Rothchild D.**, **So D.R.**, **Texier M.** and **Dean J.** (2022). The carbon footprint of machine learning training will plateau, then shrink. *Computer* **55**(7), 18–28.

Patterson D.A., **Gonzalez J.**, **Le Q.V.**, **Liang C.**, **Munguía L.-M.**, **Rothchild D.**, **So D.R.**, **Texier M.** and **Dean J.** (2021). Carbon emissions and large neural network training. ArXiv abs/2104.10350.

Ren C., **Wang D.**, **Urgaonkar B.** and **Sivasubramaniam A.** (2012). Carbon-aware energy capacity planning for datacenters. In *2012 IEEE 20th International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, pp. 391–400.

Schwartz R., **Dodge J.**, **Smith N.A.** and **Etzioni O.** (2020). Green AI. *Communications of the ACM* **63**(12), 54–63.

Sharir O., **Peleg B.** and **Shoham Y.** (2020). The cost of training NLP models: A concise overview. ArXiv abs/2004.08900.

Stewart C. and **Shen K.** (2009). Some joules are more precious than others: Managing renewable energy in the datacenter. In Proceedings of the Workshop on Power Aware Computing and Systems, pp. 15–19.

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press

Strubell E., **Ganesh A.** and **McCallum A.** (2019). *Energy and Policy Considerations for Deep Learning in NLP*. In **Korhonen A.**, **Màrquez L.** and **Traum D.** (eds), (Florence, Italy) (July): pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355. https://aclanthology.org/P19-1355/.

Wu C.-J., **Raghavendra R.**, **Gupta U.**, **Acun B.**, **Ardalani N.**, **Maeng K.**, **Chang G.**, **Aga F.**, **Huang J.**, **Bai C.**, **Gschwind M.**, **Gupta A.**, **Ott M.**, **Melnikov A.**, **Candido S.**, **Brooks D.**, **Chauhan G.**, **Lee B.**, **Lee H.-H.** and **Hazelwood K.** (2022). Sustainable AI: Environmental implications, challenges and opportunities. In Proceedings of Machine Learning and Systems, Vol. **4**, pp. 795–813.

Zhang Y., **Wang Y.** and **Wang X.** (2011). GreenWare: Greening cloud-scale data centers to maximize the use of renewable energy. In *ACM/IFIP/USENIX International Conference on Distributed Systems Platforms and open Distributed Processing*, Springer, pp. 143–164.

Zhang Y., **Wang Y.** and **Wang X.** (2012). Electricity bill capping for cloud-scale data centers that impact the power markets. In 2012 41st International Conference on Parallel Processing, pp. 440–449.

Zhao D., **Frey N.C.**, **McDonald J.**, **Hubbell M.**, **Bestor D.**, **Jones M.**, **Prout A.**, **Gadepally V.** and **Samsi S.** (2022). A green(er) world for A.I.. In 2022 IEEE International Parallel and Distributed Processing SymposiumWorkshops (IPDPSW), pp. 742–750.

Cite this article: Church K (2026). Emerging trends: Supercomputers and clouds are so last century. *Natural Language Processing* **32**, 224–231. https://doi.org/10.1017/nlp.2026.10017

https://doi.org/10.1017/nlp.2026.10017 Published online by Cambridge University Press