Peter Levine, a general partner at venture capital firm Andreessen Horowitz, has an interesting working theory. He believes that cloud computing is soon going to take a back seat to edge computing — and we will very quickly see the majority of processing taking place at the device level.

As crazy as that sounds — and he fully recognizes that it does — Levine says it's based on sound analysis of where he sees computing going — and he believes his job as an investor is to recognize where the industry is heading before it happens.

He theorizes that as devices like drones, autonomous cars and robots proliferate, they are going to require extremely rapid processing — so fast, in fact, that sending data up to the cloud and back to get an answer will simply be too slow.

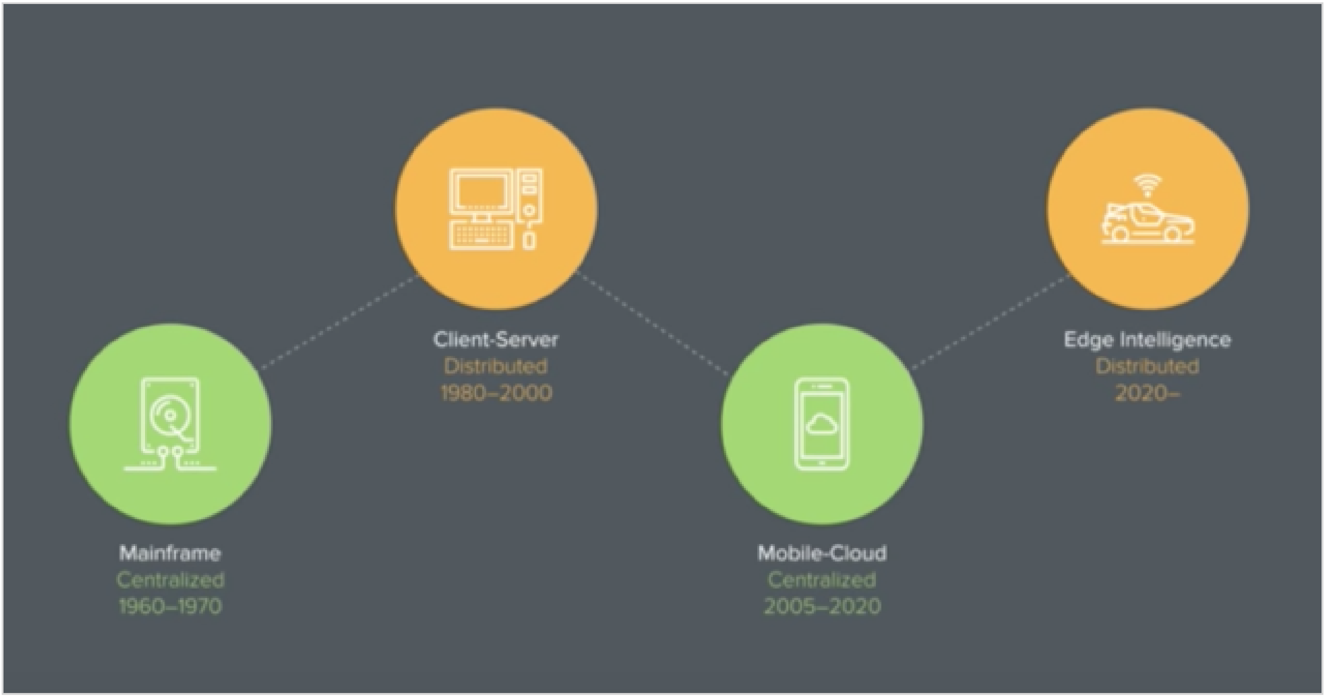
When you consider that it's taken the better part of a decade for most companies to warm to the idea of going to the cloud, Levine is saying that we are already about to supplant it and move onto the next paradigm.

That's not to say that the cloud won't continue to have a key place in the computing ecosystem. It will. But if Levine is right, its role is about to change fairly dramatically, where it will be processing data for machine learning purposes, acting as an adjunct to more immediate data processing needs.

Levine isn't alone in this thinking by any means. Other companies are beginning to recognize this, too, and we could be about to witness a massive computing shift just as we've begun to get used to the previous one.

I feel like we've been here before

If the idea of processing data at the edge sounds familiar, it should. Levine points out computing has gone in massive cycles, shifting from centralized to distributed and back again, and the coming move to the edge is just another manifestation of that.



> *Photo: Peter Levine*

In his view, it only makes sense that the next trend will swing back to a distributed system driven by the sheer volume of Internet of Things devices. When the number of devices on the planet is no longer limited by the number of humans, it has the potential to raise the number of computers in the world by an order of magnitude, and that will force a change in the way we think about computing in the future.

Levine says we are at the very beginning of this change, as we start to see the development of autonomous cars and drones, but he sees a future where this will eventually lead to the on-going proliferation of an abundance of smart devices — and it's going to happen quickly.

Processing massive amounts of data

As Levine puts it, "Think about a self-driving car, it's effectively a data center on wheels, and a drone is a data center with wings and a robot is a data center with arms and legs and a [ship] is a floating data center…" He adds, "These devices are processing vast amounts of information and that information needs to be processed in real time." What he means is that even the split-second latency required to pass information between these systems and the cloud simply takes too long.

Think about a self-driving car, it's effectively a data center on wheels.
— Peter Levine

If a car needs to a make decision, it needs the information instantly and no amount of latency is going to be acceptable.

Danielle Merfeld, VP at GE Global Research, says her company faces a similar kind of issue. GE makes huge machines like locomotives and gas turbines, generating tons of information, and they realized a few years ago, as the sensors on these massive machines generated ever-more data, it was going to require processing on the device itself at the edge, while moving only the most valuable data to the cloud for machine learning purposes.

Each machine leaves data exhaust, and if they share the best data in the cloud, and deliver it back to each individual machine, they can begin learning from one another in this virtuous cycle of data creation, processing and recirculation.

I feel the need for speed

Deepu Talla, VP and GM at Nvidia, the company that's making GPU chips that are helping fuel AI and robotics, says there are a number of reasons companies move to the edge, but it starts with a need for speed and pure practicality.

Talla says it's not just big machines that Merfeld and Levine are talking about. For some Internet of Things devices, like connected video cameras, it also ceases to be practical to send the data to the cloud just because of the pure volume involved.

As an example, he points out that there are already a half a billion connected cameras in place today with a billion expected to be deployed worldwide by 2020. As he says, once you get over 1080p quality, it really ceases to make sense to send the video to the cloud for processing, at least initially, especially if you are using the cameras in a sensitive security zone like an airport where you need to make decisions fast if there is an issue.

Then there's latency. Talla echoes Levine's thinking here, saying machines like self-driving cars and industrial robots need decisions in fractions of seconds, and there just isn't time to send the data to the cloud and back.

He adds that sometimes there are privacy issues where data could be considered too sensitive to send to the cloud and might remain on the device. Finally, companies may want to keep data at the edge because of a lack of bandwidth. If you are dealing with a location where you can't stream data, that would mean having to process it at the edge. There wouldn't be a choice.

AWS and Microsoft have noticed

AWS and Microsoft are always looking for what's coming next, so it shouldn't come as a surprise that the biggest public cloud providers have some products aimed toward the edge market already. For AWS, it's a product called Greengrass, which is providing a set of compute services directly on IoT devices when public cloud resources aren't available for whatever reason.

For Microsoft, it's Azure Stack, which offers a set of public cloud services inside a data center, giving a customer public cloud-like resources at the data center level without having to move it back and forth from the public cloud.

It's only a matter of time before we see other vendors and whole new companies begin to offer their own take on edge computing

What does it all mean?

In fact, if this change happens as Levine predicts, he thinks it's going to have a profound impact on computing as we know it. He believes it will require new ways of programming, securing and storing data, and will change how we think about machine learning. "Every area of the compute stack gets upended as we see distributed computing come back," he said. That would represent a tremendous opportunity for both startups and VCs — especially those that get in early.

And just as we saw companies ahead of the cloud and mobile curve a decade ago, Levine says he is starting to see companies planting seeds in this area. "After this video and blog series went out, we've seen companies come in, and I didn't know they existed, and they are pitching me," he told TechCrunch in an interview.

As we've seen, no form of computing ever quite goes away when a new one comes along. IBM is still selling mainframes. There are client/server networks inside many organizations across the world today and mobile/cloud will still exist, if and when Levine's vision comes to pass. But it could change how we think about computing, how we build computers and how we write programs.

Levine firmly believes that the time to start thinking about this is right now, before the change takes hold. After we are in the middle of it, the best ideas will already have been taken and it will be too late.