What's the Big Deal About Cloud Computing?
By Eric Goldscheider

The expertise of Associate Professor of Computer Science Dominique Thiébaut is much in demand in the computing world these days. He knows how to program for what has come to be known as "the cloud" and is teaching this skill to his students. His interest in how computing infrastructure is evolving in new and revolutionary ways informs his research and, in turn, helps him to guide his students in the most recent advances.

The concept of cloud computing is so new and is changing so quickly that it's difficult to define. Basically it allows the user to gain access, using the Internet, to off-site, increased computing capabilities (both storage and services) without having to maintain them. It means that users will no longer have to buy or be responsible for their own backup computers, servers and storage space or employ the workers to support them.

Consumers, maybe even without realizing it, are already using the cloud every time they check their e-mail, listen to music or watch movies online.

The cloud concept, says Thiébaut, has tremendous implications for the consumer market, as applications such as Netflix, Apple's Siri (a voice-response system) and other customized data back-up systems take off. It also portends momentous changes in areas like scientific research, where deploying powerful computing power more logically will make it accessible to more investigators.

The cloud also has huge implications for research. An investigator at an institution like Smith can think big, very big in fact, when devising experiments

that heretofore only institutions with large computing capacity could afford. And that's just one small example of the implications of making large-scale computing more widely accessible.

One of Thiébaut's research interests has led him to enlist students in figuring out approaches to conducting statistical analyses of Wikipedia. At the time, in mid-2010, Wikipedia's English language pages held about 37 gigabytes of information. One of the challenges was finding enough computers on campus that, if harnessed collectively, could "process" this information. In this case they wanted to count discreet words and two-word combinations as a way of yielding a metric with which to analyze what a site might be saying about the broader culture. (For what it's worth, "United States" and "New York" are both in the top-100 two-word combinations.) Thiébaut is also enthused by graphic displays of such things as the history of Wikipedia's page on evolution, where the tracked changes highlight cultural struggles, as illustrated by creationist versus scientific approaches.

At one level "the cloud" is simply about massing huge amounts of computing power under a central game plan; for example, buildings used for cloud computing (which are often constructed near power plants) house racks and racks of computers the size of pizza boxes. Each computer is constantly communicating with counterparts around the room and around the world in ways that maximize their collective power. Using the Internet, clients tap in for a fee, paying only for the amount of computing they need--just as they would pay for the electricity they use.

It's an infrastructure that enables a brand new business model, and hence a potential surge in start-ups, and drastically reduces the costs of IT. You need a thousand computers for four hours tomorrow afternoon? No problem. Current costs, according to Thiébaut, are between 10 and 50 cents an hour per computer and about 10 cents a terabyte per month for storage.

But how is all that information sustained? That's where the special challenges of programming for the cloud come in. "When you have a farm of computers, a whole building," Thiébaut explains, "you know that at any moment one or two are dying. They just fail." This poses an obvious conundrum for programmers who can't afford to have a piece of their puzzle fall out at any moment. "Google is the one that understood that best," says Thiébaut, so they set out to create an infrastructure "where failure is recognized and even expected."

The most useful analogy for understanding how this is accomplished is by comparing the separate computing tasks to "bubbles," says Thiébaut. The tasks are encapsulated in tuples (an ordered set of data) that are launched, as you might a spray of bubbles, into the computing environment. "Then you are going to have all these different layers that grab these bubbles and somehow consolidate them, kill some that are not necessary, and at the end you have a reducing step where you get the result that you want, which is a huge bubble of the data you are interested in," said Thiébaut,

To understand the speed inherent in this centralized yet at the same time diffused approach to computing, you need look no further than Google's ability to guess the word you are thinking of before you've finished typing it.

Because the cloud process is evolving so rapidly and so many competing interests are working on it, the approaches and the language used to describe them are not uniform—despite the common goal. The three big players today in cloud computing are Google, Microsoft and Amazon, and each is developing its own computing culture with specialized vocabularies. One of Thiébaut's former students works for Google and, he explains, even if she were allowed to talk to him about her projects (which she is not) he might not be able to understand her.

These inconsistencies were apparent this past May at the First International

Conference on Cloud Computing and Service Science, held in Noordwijkerhout, Holland, where Thiébaut presented a paper on his Wikipedia analysis. "Everyone is doing things differently," he says, "… and sometimes when they are talking about the same thing, they have a different vocabulary."

Other companies, notably Facebook and Apple, have their own clouds, which are not available for hire. Of the big three, Amazon is the favorite of programmers, according to Thiébaut, even though you need to know how to talk to it in symbols and numbers. The other two in the race for dominance as so-called "super clouds" have interfaces, such as the Windows environment, that are evolving to fit the needs of their clients. Many are intermediaries selling services such as back-up systems, intelligent music delivery and all manner of data analyses that advertisers crave, such as demographic information and consumer buying trends.

Consumers are seeing the cloud in a ballooning number of places. If you have an iPhone with Siri—the voice inside the mind of the machine—you are floating on Apple's exclusive cloud. "It will even understand my French accent," says Thiébaut, marveling at the device's ability to answer, almost immediately, such questions as, what is the weather? or, where is the nearest Mexican food? Car manufacturers are connecting diagnostic tools to the cloud, distributors use the cloud to track inventory, and innovators like Pandora.com have created a cloud-based application that facilitates a highly personalized and almost instantaneous stream of music.

Thiébaut's interests are leading him in the direction of figuring out ways of analyzing what he calls "data hoses," such as Twitter and Facebook, which every day produce torrents of clues about the pulse of cyberspace. He agrees that the term "cultural data mining" aptly describes this endeavor.

Of late "the cloud…has become a buzz word. If you don't say 'cloud' it means

that you are not right in the wave…," says Thiébaut. But at the same time, "It is in its infancy, it's exciting, and there's plenty to do."

SIDEBAR

When the *New York Times* decided to digitize its public domain archives four years ago they turned to Amazon for less than a full weekend's worth of computing time to translate every article and every advertisement from 1851 to 1922 into zeros and ones that can be manipulated at will. The cost would have been prohibitive if they had had to purchase the computing infrastructure for such a project.

Similarly, if you want to sell chocolates over the Internet and your business goes viral on Valentine's Day and then goes into a slump until Easter and Mother's Day, you can dial up and dial down the amount of computing you need almost to the minute.

"In the past a small company needed roughly $100,000 for a serious computer infrastructure," says Thiébaut. That didn't include the price of technicians to keep it on track or the cost of updating obsolete equipment. Today, a burgeoning array of intermediaries offer "cloud solutions" that let such an enterprise key its computing costs almost organically to the ebb and flow of the demand for its product, and they avoid sullying their brand by having slews of disappointed potential customers due to a system overload or crash.

The cloud is an apt analogy for this type of data storage and management, says Thiébaut. "It's way up there, it's all around, and you don't really know where it is," he said. "You never reach there, you never go there, it's hidden, it's far away, it's behind layers and it can be very big."

END