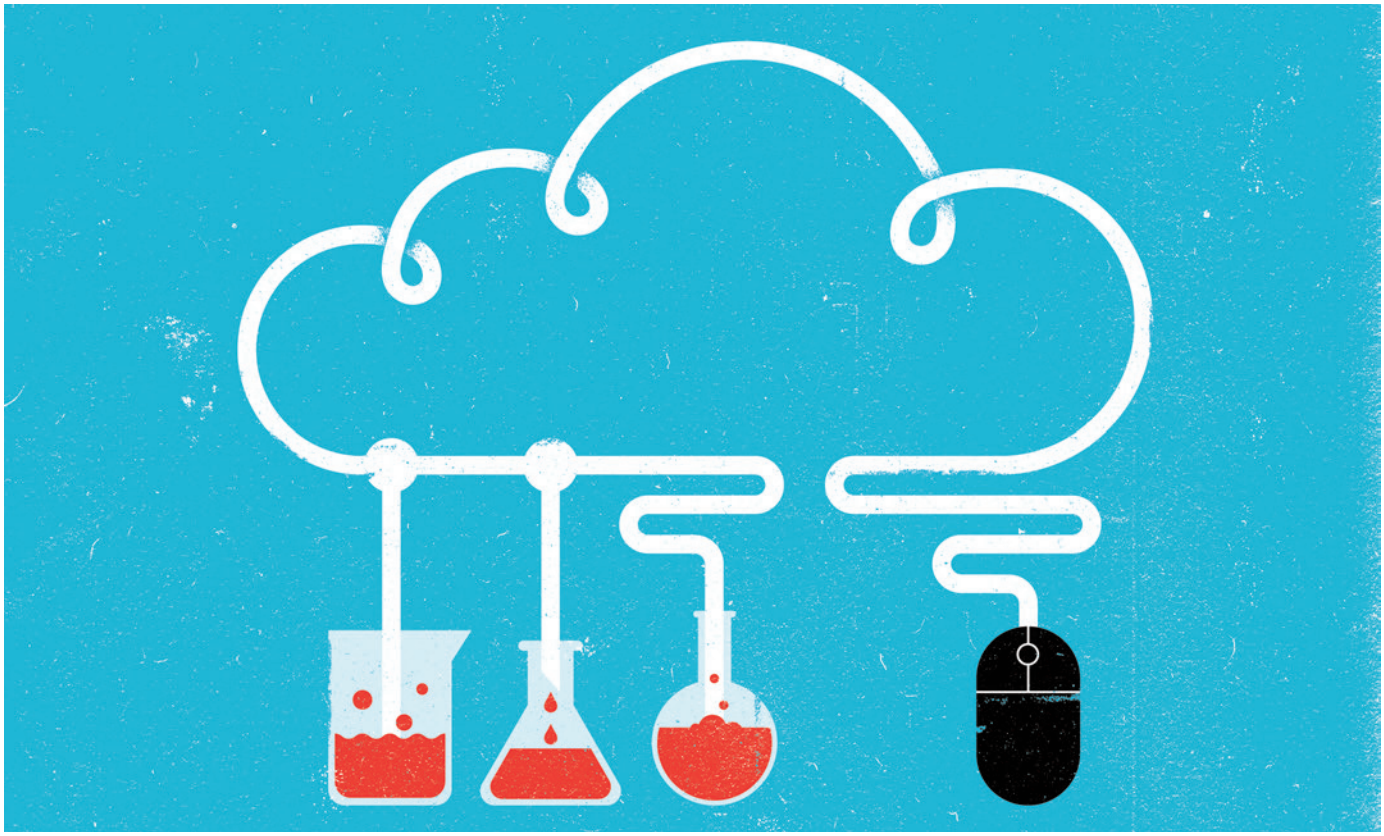
TOOLBOX

HOW TO CATCH A CLOUD

Why cloud computing is attracting scientists — and advice from experienced researchers on how to get started.



THE PROJECT TWINS

BY NADIA DRAKE

In February, computer scientist Mark Howison was preparing to analyse RNA extracted from two dozen siphonophores — marine animals closely related to jellyfish and coral. But the local high-performance computer at Brown University in Providence, Rhode Island, was not back up to full reliability after maintenance. So Howison fired up Amazon's Elastic Compute Cloud and bid on a few 'spot instances' — vacant computing capacity that Amazon offers to bidders at a discounted price. After about two hours of fiddling, he had configured a virtual machine to run his software, and had uploaded the siphonophore sequences. Fourteen hours and US$61 later, the analysis was done.

Researchers such as Howison are increasingly renting computing resources over the Internet from commercial providers such as Amazon, Google and Microsoft — and not just for emergency backup. As noted in a 2013 report sponsored by the US National Science Foundation (NSF) in Arlington, Virginia, the cloud provides labs with access to computing capabilities that they might not otherwise have (see go.nature.com/mxh4xy). Scientists who need bursts of computing power — such as seismologists combing through data from sensors after an earthquake or astronomers processing observations from space telescopes — can rent extra capacity as needed, instead of paying for permanent hardware.

Scientists can configure their cloud environment to suit their requirements. Although cloud computing cannot handle analyses that require a state-of-the-art supercomputer or quick communication between machines, it may be just right for projects that are too big to tackle on a desktop, but too small to merit a high-performance supercomputer. And working online makes it easy for teams to collaborate by sharing virtual snapshots of their data, software and computing configuration.

But shifting science into the cloud is not a trivial task. "You need a technical background. It's not really designed for an end user like a scientist," says Howison. Although the activation energy might be high, there are recommended routes for scientists who want to try setting up a cloud environment for their own research group or lab.

A DIY GUIDE TO CLOUD COMPUTING

Most cloud platforms require users to have some basic computing skills, such as an understanding of how to work in the command line, and a familiarity with operating systems and file structures. Once researchers have a strong foundation, the next step is to try working in a cloud.

The most user-friendly cloud for scientists, says plant biologist Andreas Madlung, could be the platform Atmosphere, which was created as part of a collaborative cyber ▶

© 2015 Macmillan Publishers Limited. All rights reserved

infrastructure project called iPlant. Funded by the NSF and led by three US universities and the Cold Spring Harbor Laboratory in Long Island, New York, iPlant has been helping scientists to share software and run free analyses in the cloud since 2008.

Designed with scientists in mind, the platform's interface comes with pre-loaded software, a suite of practice data sets and discussion forums for users to help each other to tackle problems. Madlung, at the University of Puget Sound in Tacoma, Washington, teaches an undergraduate bioinformatics course that includes a section on cloud computing. He first introduces his students to the Unix operating system, then has them use that knowledge to analyse RNA sequence data on Atmosphere.

Those who sign up with iPlant are automatically given what equates to around 168 hours of processing time a month, and can request more if needed. Users can load up virtual machines with any extra software that they need, and if a job is too much for standard equipment to handle, tasks can be offloaded to a supercomputer at the Texas Advanced Computing Center in Austin, where iPlant has a guaranteed allocation.

Biologist Mike Covington of the University of California, Davis, shifted his lab's computing work to iPlant after its servers kept crashing because they were overloaded. He has also made copies ('images') of his own virtual machine, so that his collaborators — and any iPlant user — can log in and access the same software, data and computing configuration. "If I spend several hours setting up my virtual machine perfectly for *de novo* genome assembly [reconstructing full-length sequences from short fragments of DNA], I can quickly and easily make it available to any other scientist in the world that wants to do *de novo* assembly with their own data," Covington says.

Such virtual snapshots may become standard for projects that require computational work. Anyone who wants to reproduce, for example, the microbial-genome analysis described in one paper can access a snapshot of the authors' virtual machine on the Amazon cloud, simply by paying for Amazon computing time (B. Ragan-Kelley *et al. ISME J.* **7,** 461–464; 2013).

PICK A CLOUD

For some researchers, choosing a cloud is straightforward. Scientists at CERN, Europe's particle-physics laboratory near Geneva, Switzerland, have had access to a massive internal cloud running on the software platform OpenStack since 2013. A handful of institutions, such as Cornell University in New York and the University of Notre Dame in Indiana, have developed computing clouds, too. Some, including Notre Dame, outsource their clouds to companies such as Rackspace Private Cloud, a multi-national firm in San Antonio, Texas, that sets up and manages cloud services for users. But for scientists who are not at an institution with a fully functional campus cloud, bushwhacking through the jungle of cloud options can be a frustrating adventure (see 'A guide for the perplexed'). Cloud system set-up can vary, and proficiency with one provider does not guarantee an easy transition to others.

Casey Dunn, an evolutionary biologist who works with Howison at Brown University, prefers to train students on commercial platforms. "When they go on to a postdoc somewhere else or start their own lab, they'll still be able to log into Amazon," he says.

Somalee Datta, the director of bioinformatics at Stanford University's Center for Genomics and Personalized Medicine in California, is using Google's cloud platform to support the centre's enormous amount of genomics data and computing demand, rather than relying only on the servers available at Stanford. She chose Google, she says, for several reasons: the company's developers were actively making tools available for genomics researchers, Google had demonstrated interest in health-care research — and the price was right.

CLOUD CONCERNS

For Datta and others, one key issue surrounding cloud computing is security. "It's a big concern," she says. "Hackers understand where the value is, and they will turn their attention towards that." Still, Datta thinks that clouds are no more or less secure than any other computer network. A university cloud system, for example, is only as solid as the university's firewall. "If I were working on my own or at a small college or company, I would probably feel more secure with Google's cloud," Datta says (although Stanford has its own army of engineers watching security). The truth is, anyone working with extremely sensitive data might be better off keeping it away from the Internet altogether.

Another key issue for researchers who are venturing into cloud computing is the level of tech support needed. Getting software to run on a new system can take days, and determining how much computing power or memory a virtual machine needs can be an exercise in trial and error. All cloud providers offer training and tutorials, but dedicated support staff are more commonly found at universities with campus clouds.

Despite the challenges, cloud computing is increasingly appealing to scientists, says Darrin Hanson, vice-president of Rackspace Private Cloud. "The last few years have been mostly people who are absolutely out on the bleeding edge," he says. "But now we're starting to see an influx of adopters."

That isn't too surprising, Dunn says — the cloud is not as foreign as it can sometimes sound. "Nearly all consumer computer products now have a cloud component, be it mobile apps, content-streaming services like Netflix or desktop tools like Dropbox," he says. "Research computing is not on the vanguard of some crazy and risky unknown frontier — we are just undergoing the same transitions that are already well under way in industry and the consumer marketplace." ■

Nadia Drake *is a freelance science writer in San Francisco, California.*

CLOUD RESOURCES

A guide for the perplexed

- **Clouds for researchers:** The largest commercial providers include **Amazon's Elastic Compute Cloud**, **Microsoft's Azure** and **Google's Cloud Platform.** Other services are **Terminal.com**, aimed specifically at research; the (free) **Atmosphere** cloud platform, from the US National Science Foundation-backed iPlant collaboration; **SageMathCloud; Cornell University's RedCloud**; **Digital Ocean** — known for quick deployment of cloud apps; and **Rackspace** — a company that sets up clouds using **OpenStack**, an open-source cloud-software platform that the firm developed jointly with NASA.

- **Useful resources for cloud explorers**: **StarCluster** is a tool developed at the Massachusetts Institute of Technology in Cambridge that helps to build a virtual research-computing cluster on Amazon's platform. **Docker** is an open-source platform that allows researchers to share a snapshot of the code, computing environment and data used to generate analyses. **Project Jupyter** are shareable notebooks that make data, code and analysis easily accessible — and interactive (H. Shen, *Nature* **515,** 151–152; 2014). **Nimbus**, partly developed by the Argonne National Laboratory in Illinois, helps to turn a normal computing cluster into a cloud system accessible by remote users.

- **Other computing resources:** *Practical Computing for Biologists*, by Casey Dunn and Steven Haddock (Palgrave Macmillan; 2011).
The **Software Carpentry** computing workshops (see go.nature.com/jg86jj).
The **University of Washington's eScience Institute** advice on "Which compute platform should I use"? (See go.nature.com/iazoio).

Links to these resources, including tutorials, are available at the online version of this article. N.D.

© 2015 Macmillan Publishers Limited. All rights reserved