COMPUTING

# Research brings cloud computing costs back to earth

STORY BY Mandy Thoo

KEY POINTS

- **The search is on to cut the costs of data storage on web-based and remote servers**
- **Savings for heavy users of cloud computing – science, government and industry – are potentially huge**
- **Swinburne researchers have set out to change the way users manage remote databanks**

**RESEARCHERS ARE LOOKING** for ways to reduce the high cost of internet data storage and retrieval in cloud computing – running software and managing data on a remote computer, rather than your own.

Social media such as Facebook and Flickr are simple examples of cloud computing, but the drain on resources from these doesn't compare to the volumes of high-end data generated by the world's research institutions, healthcare systems and industries.

Government agencies such as the Australian Taxation Office, Bureau of Statistics, and Treasury are examples of heavy users of cloud computing services, and the costs to them and others are high and rising. A new, more cost-effective model is needed for heavy users of cloud computing and this is the task of a team at Swinburne University of Technology.

As Professor Yun Yang from the Swinburne University Centre for Computing and Engineering Software Systems (SUCCESS) explains, cloud computing offers almost unlimited space for data storage and processing, but the current usage charges mean the costs are expanding at the same near-limitless rate.

### Savings to be made

"Users have to pay for data storage, computation and data transfer for these pay-as-you-go services. Also, unnecessary data storage means greater electricity consumption, it is less environmentally friendly, lessening the benefits of using cloud computing.

"In his report to the federal government, Sir Peter Gershon estimated $1 billion can be saved if the Australian government develops a data centre strategy – the core for cloud computing – for the next 15 years."

Professor Yang, who is working with colleagues Professor John Grundy and Dr Jinjun Chen, illustrates the issue facing cloud computing when he notes that scientific fields such as astronomy, high-energy physics, bioinformatics and medical imaging technology can generate gigabytes of data per second.

Professor Grundy, director of SUCCESS, also points out that a complication for the Swinburne team is that researchers store two kinds of data: raw data and the intermediate data generated from processing this initial data. Raw data must be securely stored because it cannot be regenerated. The main catch, Professor Grundy says, lies in storing intermediate data, and this gets expensive. But if the user deletes all the intermediate datasets and has to regenerate them later through further computation, the costs can be even higher.

"The trade-off is going to be between storage cost and computation cost. Finding this balance is complex, and there are currently no decision-making tools to advise on whether to store or delete intermediate datasets, and if to store, which ones," Professor Grundy says.

### Cloud cost calculators

With funding from an Australian Research Council Discovery Project Grant, the researchers have developed two potential strategies to make cloud computing more affordable.

In the first, they have developed a mathematical model that lets users calculate the minimum storage cost. The formula factors in the size of the initial datasets, the rates charged by the service provider and the amount of intermediate data stored in the specified time. "The formula can be used to find the best deals for storing data in the cloud," Professor Yang says.

The second proposal, the Intermediate Data-dependency Graph (IDG), answers whether to spend on storage or computation for intermediate datasets.

"IDG records how each intermediate dataset is generated from the ones before it. It shows the generation relationships of the datasets." Or in other words, how A leads to B, then C, and then to D. "The (data centre) strategy can then decide which ones to delete."

If deleted intermediate datasets need to be regenerated, the system wouldn't have to start at the original data. Instead, guided by the IDG, the system could find the nearest predecessors of the datasets. If the process has been A, B, C, D and C is removed, the model would revert back to B for computation, rather than A. "This can save computation cost, time and electricity consumption," Professor Grundy says.

The researchers have been evaluating the two solutions in tandem by simulating a pulsar survey used to crunch information from radio telescopes.

"Searching for pulsars – rapidly spinning stars that beam light – is a typical scientific application," Professor Yang says. "It generates vast amounts of data – typically at one gigabyte per second. That data will be processed and may be reanalysed by astronomers all over the world for years to come.

"We used the prices offered by Amazon cloud's cost model for this evaluation. For example, 15 cents per gigabyte per month for storage, and 10 cents per hour for computation."

From one set of raw beam data collected by the telescope, the pulsar application generated six intermediate datasets. The model generated three different cost scenarios. The minimum cost for one hour of observation data from the telescope and storing intermediate data for 30 days was $200; for storing no data and regenerating when needed, $1000; and for storing all intermediate data, $390.

This gave the researchers options for which data to keep, and which to delete. "We could delete the intermediate datasets that were large in size but with lower generation expenses, and save the ones that were costly to generate, even though small in size," Professor Yang says.

These are only a few of the solutions the researchers have come up with so far. To cater to different sectors, the group is also working on models that will allow users to determine the minimum cost on-the-fly, and as frequently as they wish.



Bringing data costs down to earth: Professors John Grundy and Yun Yang.