Resource amplification through elastic computing

Complex processing of real-world data is only possible with enormous computing resources. We can obtain these resources through the novel property of *elasticity*. Elasticity is the ability to quickly ramp up and down resource usage on demand. In today's computing infrastructure, elasticity is provided primarily by *cloud computing*, hosted in data centers. Applications that process complex data can absorb very great computing and storage resources. This can be not only because of the problem complexity (e.g., NP-completeness), but simply because of the sheer amount of work needed to store and process large quantities of data.

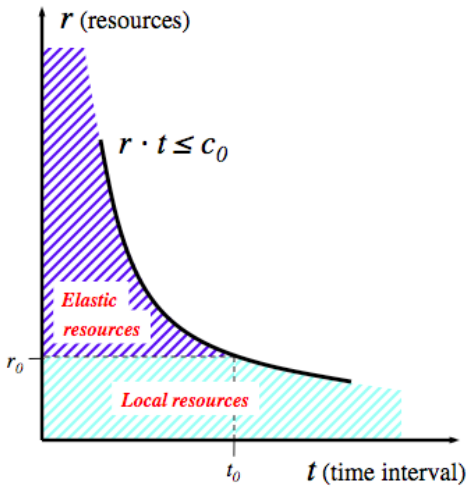

Figure 1: Additional resources available because of elasticity

By using elasticity, enormous resources can be made available to applications for short time periods. This is economical because these resources are amortized among all applications and their users. On a cloud, the run-time cost of an application is proportional in a first approximation to the product of the amount of resources (computing and storage) and the time that they are needed. Figure 1 shows what this implies. For an application whose resources have total cost c_0, the lightly shaded area shows the different combinations of resources r and time interval t that are possible when running on a single computer. The resources are limited by r_0, which is the maximum that the computer can provide. If the application runs on a cloud, then the darkly shaded area becomes available: the cloud can provide enormously more resources with cost limited to c_0 as long as the resources r are used for a limited time t, according to the formula $r \leq c_0/t$. The darkly shaded area is the focus of the Flagship. In principle, the amount of resources that are available is unlimited, if they are needed for a sufficiently short time. In practice, the maximum is limited by what the cloud is able to provide according to the application's service contract. In today's clouds this can go up to about one thousand with the appropriate contract. In tomorrow's elastic infrastructures, the limit should be determined by what applications need.

The ability to increase resource usage for the same cost, as long as the resources are used for just a short time, is very important for real-world computing applications. They will need many resources both in the learning and query phases (as explained in the next section). Elastic computing has the potential to provide these resources practically and economically. However, there are many practical limitations of current elastic infrastructures (primarily clouds): they are slow in ramping up and down resources according to need, they are limited in how far they can ramp up (taking minutes instead of, e.g., fractions of a second to ramp up), and they have other problems (e.g., security, fault tolerance, bandwidth, latency). These are strong limitations that will take many years to overcome. That is why it is important to start working on them right away.

Elastic computing: clouds and beyond

Real-world computing will succeed or fail depending on the abilities of the elastic infrastructure that supports it. In this section we briefly recapitulate the state of the art for elastic computing and define its main properties. Today's main source of elastic computing is *cloud platforms*. A cloud is a form of client/server with novel properties that derive from its large scale. Cloud computing uses the memory and processing power of a large number of computing nodes gathered together in facilities called data centers and linked through high performance networks. Cloud users have at their disposal considerable computing resources that are both flexible and modestly priced. For example, many Web applications execute on a cloud instead of on client machines. Cloud computing has three properties that distinguish it from other forms of client/server computing [EGR2010]:

1. *Virtualization*: The ability to run applications in customized environments that are insulated from the underlying hardware. Virtualization greatly simplifies software installation and maintenance. Current virtualization techniques have a performance penalty, but they are still practical for many applications such as enterprise computing and support for small, networked devices such as mobile phones.
2. *Scalability*: The ability to provide almost any amount of computing and storage resources. This is possible because of the large size of the data center. Because of their size, data centers have an economy of scale. The Berkeley report measures them as five to seven times cheaper than enterprise installations [ARM2009].
3. *Elasticity*: The ability of an application to quickly ramp up and down resource usage on demand. Because of this ability, cloud usage is metered: the user pays only for what is actually used, with almost no entry threshold. Furthermore, the actual cost of resources is low because of the economy of scale of the data center and the amortization of its cost over all users.

Of these three properties, the true game changer is *elasticity*. It enables a whole new class of applications that were not possible before, namely applications that need large computing and storage resources for short time periods [VAN2010]. Previously such applications could not be run since the resources were simply not available. On a cloud, the resources can be requested quickly through the elastic computing mechanisms, and released when they are no longer needed (up to the practical limits of the cloud implementation).

Clouds are currently hosted in data centers, but this will quickly become inadequate to meet the demands for real-world computing. Fortunately, Internet resources outside of data centers dwarf the largest clouds by three orders of magnitude: in Jan. 2010 there were 800,000,000 Internet hosts [ISC2010] versus less than 1,000,000 hosts in the biggest cloud [COS2009]. We therefore predict that most Internet hosts will eventually acquire cloud-like abilities (the three properties mentioned above) and be federated into peer-to-peer clouds, to support the Internet's ever-increasing appetite for data-intensive applications. Medium- and small-sized clouds requiring modest investments will complete the picture. One of the goals of the Flagship is to catalyze this transition to a fully elastic Internet.

References

[ARM2009] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy H. Katz, Andy Konwinski, Gunho Lee, David A. Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. “Above the Clouds: A Berkeley View of Cloud Computing,” UC Berkeley, Technical Report UCB/EECS- 2009-28, Feb. 10, 2009.

[COS2009] Paolo Costa. “The Hitchhiker’s Guide to the Data Centers Galaxy,” Microsoft Research Cambridge, 2009.

[EGR2010] European Commission Expert Group Report. *The Future of Cloud Computing: Opportunies for European Cloud Computing Beyond 2010*. Version 1.0. Eds.: Keith Jeffery, Burkhard Neidecker-Lutz. Rapporteur: Lutz Schubert. Jan. 2010.

[ISC2010] Internet Systems Consortium, Inc. “ISC Domain Survey,” www.isc.org, 2010.

[VAN2010] Peter Van Roy. “Scale and Design for Peer-to-Peer and Cloud,” Invited talk at TTI Vanguard Conference *Matters of Scale*, July 20-21, 2010, London, UK. See www.info.ucl.ac.be/~pvr/TTIVanguardPVR.pdf.