Cloud why now?

By Dave Levy, Cithub Ltd., London

How new is Cloud Computing? It is a clear evolution of two trends in IT architecture that have had immense and limited success. The successful trend is distributed computing the less successful trend utility computing. What is driving this evolution and why now? This article has a quick look at the trends that have brought us to this point and looks at the fact that like most economic revolutions, it's a confluence of socio-economics and science. The article concludes by looking at the paradox of data, to see how it is both a driver of change, and an inhibitor.

The internet was first conceived as a network of computers and it is evolving into a network of people, and will move on to become a network of things. This evolution changes the notion of IT scale. Modern IT enterprise systems, attached to the internet has empowered customers, permitted mass customisation and in some industries[1] even permitted the delivery of product. This all changes the definition of macro-economic scale; it is no longer bound by the US economy. The internet and ERP are creating new markets in a new scale.

Current and traditional IT architectures for consumer and business IT are horrendously wasteful, with utilisations often in single digits, and average utilisations more often in the 10% - 15% range. This is an opportunity for micro economic competitive advantage and at the macro- level plays into both the economic efficiency and the Green agenda, of which one part is the about the manufacturing efficiency of the IT vendors. However the greatest inefficiency is often the actual utilisations of IT equipment. New architectures are designed to make today's enterprise IT more efficient by solving the problems of low utilisation. Massive scale exacerbates the under utilisation problem; one might be able to afford 10% utilisation if one has 10's of computers, but if one has 1000's or 10s of thousands, the waste is unsustainable.

What are the technology shifts causing the movement to new ways of doing things? The most significant changes stem from the ending of Moore's Law. It's too hard & expensive to continue to build more powerful CPUs and all CPU architects are looking at multi-threading/multi-core solutions. This might be seen as shrinking the server onto a chip, but it is also beginning to break the fundamentals of Symmetric Multi-Processing[2]. Once programmers, be they operating system authors or applications engineers need to take account of multiple systems then they can ask "Why stop at two?" Distributed computing and post SMP systems architectures have a different speed relationship between hierarchical memory structures, CPU and network speeds. Both the science of CPU architecture and the new dimensions of scale are pushing IT architecture towards a new, scale out distributed platform; architects can no longer shrink many application into a single computer. The final piece of the jig saw is the improving technology of networking itself. We have seen massive improvements in speed and cost that enable the connection of people, computers and even appliances and tags.

The problem of IT scale has three dimensions, the number of users, the amount of data, the complexity of the algorithm i.e. the number of cycles. Two key software innovations that enable these new massively scalable architectures, are the Google File System and its imitators and inspirations and Google's Map/Reduce both of which were created to allow the parallelisation of 'Search', but innovators have begun to improve and apply the technology and algorithms in new (& old) ways. These are both attempts to remove seriality and parallelise the infrastructure to permit new distributed platforms.

The two key scalable paradigms today are HPC and Social Network sites. They have different

---

1 Financial Services and Entertainment & News Media
2 SMP systems have several or many CPUs sharing access to a single memory address space, at a uniform speed.

characteristics which I have summed up as

> *"Where you run one application on a distributed computing platform you have high performance computing (or grid computing) and where you run many copies of one application on a distributed computing you have web 2.0 computing"*

although Web 2.0 or modern massive e-commerce/portal/social networking sites rarely have only one application, they do have a limited set of applications, the management of, and relevance of which eclipses the functional requirements and architectural requirements of the rest of the enterprise's IT. This gives us a good architectural definition between Grid & Cloud. The third piece of the puzzle is

> *"...where you have many applications running on a single distributed computing platform you have "Cloud Computing".*

The reason we are seeing this innovation and change today is because new dimensions of economic and IT scale and new science requires a new architecture. These new distributed computing platforms can be optimised in various ways, for high performance computing, or for massive user scale. Many scientists have been building grids, mainly for modelling applications, be it pure science, engineering or financial analysis; and these have been joined in the developed economy's telcos (and amazon) where a number of mass scale portals have been built.

Economies of Scale and the ability of large providers to more easily deal with peaks & troughs will lead to economic consolidation. Outsourcing of IT supply requires new models of metering & billing. Billing and metering require solving the problems of actual resource utilisation, the aggregation of usage across the platform, the attribution of cost to the usage and the billing of the customer. It is in solving these problems, that today's architects need to borrow from the past designs of the late '90s utility computing providers. These problems should not be too hard to solve and it'll be interesting to see if the need for a common billing solution will drive integration across the hardware component classes, (servers, storage & networks) and hence the operating systems. The billing and metering solutions need to be transparent as well. Customers need to be able to check their bills against their delivery notes, or whatever the virtual equivalents become. The building blocks are in place and as virtualisation technology improves, usage capture will become easier.

Once the applications architects have worked out how to split the work up into pieces that can be run in parallel, then the infrastructure architects can begin to work out how to get multiple installations/clusters to work together and see if we can permit multiple installations to perform this distributed workload. This may even enable a spot market in cloud computing, and it'll be interesting to see if the different computer architectures will become a sort of foreign exchange market.

A final though is that while the explosive growth in data is causing these new architectures, its very volume and time cost to move it to the computers is becoming prohibitive. Today, data needs to be close to its processing computers i.e. the CPUs, and this can be seen in the design of the modern HPC cluster built in the USA, Japan and Europe, where the amount of storage is increasing. The 'data latency' problem doesn't inhibit the building of 'cloud computing', if anything it enables it, but it does create a mighty, if not prohibitive cost of to move between providers. This new lock-in may lead to a much slower adoption than should otherwise occur and this will minimise the macro- benefits. The other data related inhibitor is that of security and regulatory compliance. This is becoming better understood these days; but the bottom line is that regulators may require that their agents can visit the disks and that the data users can guarantee that they can meet their privacy commitments to their data subjects and the regulators. It would seem that the over-riding principal is that data can't move from a high-privacy rights jurisdiction to a lower privacy rights one.

The growth of the internet and the ending of Moore's Law have created a point at which IT

scale need to change up a gear and new architectures based on distributed/co-operative computing are required. The architecture of co-operation, based on academia's parallel computing solutions has permitted the construction of the tera-architecture sites (Google/Microsoft/Facebook). Standard solutions are still required for metering and billing, and also for allowing specific managed clusters to work in collaboration, but these problems are likely to be solved. Since these changes are driven by new IT scale, the coming development of the "internet of things" will only make this worse.

ooOOOoo

Dave Levy
Dec 2009, London

Citihub, Smart Infrastructure Solutions –
London – New York - Singapore