**Cloud Value Proposition for Science and Engineering**

Clouds offer improved functionality and better cost-performance than traditional approaches in many areas of computational science and engineering. Many of these opportunities have not been explored in depth as there is currently no viable business model as clouds charged as operating funds (bearing overhead) must compete with no-cost resources available through universities and federal initiatives. On general principles one can expect clouds to be the most economical computing resource as they offer economies of scale (one has around 100,000 servers in a large cloud data center) and their internet access model can allow cloud centers to be placed in optimal locations where operating costs are low and environmental impact is minimal. Of course current national supercomputer resources operate near 100% utilization (whereas clouds typically operate below full utilization allowing an attractive interactive model) and often are directly or indirectly subsidized by the host organization and this obscures the comparison of cloud and traditional scientific computing approaches.

Clouds offer interesting opportunities as both infrastructure (IaaS) and software (PaaS) levels. Their software model has been developed for the largest scale data intensive applications in the e-commerce, social media and search arenas. These have been reinforced by the commercial cloud focus as general next generation enterprise data center technology.

Comparing clouds, grids (distributed systems) and supercomputers, clouds have synchronization and communication costs that lie between those of distributed systems and supercomputers. Further clouds tend to be optimized for external access and not for inter node communication performance. Thus highly parallel large scale simulations are not likely to move to clouds in the near future and should remain staple of traditional supercomputers. However there are two important classes of applications where clouds could perform well and offer attractive cost-performance, interactive elastic (on demand) use and powerful new software platforms. These classes are

1. Pleasingly parallel applications and with some overlap
2. Data-intensive applications

Clouds offer an interesting high throughput computing model for the pleasingly parallel case where there are two important cases – namely parallelism over users and usages. The former is illustrated by the many users of a Web 2.0 site in commercial applications and by support of the “long tail of science” (the many small users with individual jobs) in scientific case. The success of the European Venus-C project on the Azure cloud is a good example here. Parallelism of usages could be illustrated by particle physics data analysis (each event set can be analyzed independently) or the support of Sensor nets or more generally the “Internet of Things” where 24 billion devices are predicted on the Internet by 2020; each sensor is naturally connected elastically (as individual sensors such as smart phones do not have 100% duty cycle) to a core in the cloud.

The data intensive application is perhaps deeper as true parallelism is often needed and new programming models have been developed such as MapReduce and its various extensions. Simulations need communication which is often point to point and with small messages requiring low latency. Data intensive applications typically do not have this pattern. They require data-compute co-location and use “reductions” (such as global sums) and “broadcasts” with large messages that require optimizations seen in some MapReduce systems that are different from traditional MPI. For example, internet search is probably the world’s largest scale data analysis application and cloud architectures are highly suited to this. However clustering (and more generally expectation maximization) and core statistics such as histogramming are well supported by clouds.