Finding Tropical Cyclones On Clouds

Daren Hasenkamp (dhasenkamp@berkeley.edu)
University of California at Berkeley, Lawrence Berkeley National laboratory

Summary: This work uses virtualization-based cloud computing to analyze trends of tropical cyclones in climate simulation data. Virtualization is attractive here because it can provide an environment familiar to climatologists and their analysis tools. We created virtual machines (VMs) and ran them on the Magellan Scientific Cloud at Argonne National Laboratory. Our VM communicates with instances of itself to split up and analyze large datasets in parallel. In a preliminary test, we used this virtual climate analysis platform to analyze ~500GB of climate data. Using 34 VMs, the total analysis time was reduced by a factor of ~40 from traditional personal workstation-based analysis. The main advantages of our method are that the level of parallelism is easily configurable, and software dependency resolution is simple. This initial work demonstrates that a cloud computing system is a viable platform for distributed scientific data analysis traditionally conducted on dedicated supercomputing systems.

Problem: Elevated levels of carbon dioxide in the atmosphere could significantly affect human life. To predict the extent of the impact, climatologists use computer models to simulate the global climate many years into the future [3,4], producing petabytes of output. They must analyze this output with sophisticated programs, which contain numerous library dependencies and are often difficult to run outside the authors' workstation. Using a typical workstation, analysis jobs can take days or weeks, and moving analysis onto supercomputers is often difficult due to software dependencies. Hence, a computational platform coupling high compute capacity with an environment climatologists can customize to their liking is desirable. This work seeks to create such a platform by utilizing the computing power available through the cloud computing paradigm [1,2]. We used this computing power to find tropical storms in climate simulation output using a program called TSTORMS.

Cloud computing and Our VM: Cloud computing systems provide an abstraction for programmers to use powerful compute resources in a controllable environment by allowing users to create virtualized environments to run jobs in. In this study, we used the Magellen system from Argonne Leadership Computing Facility [7], which uses the Eucalyptus virtualization platform [5]. We created a VM to perform parallel analysis of climate simulation data. Each VM instance analyzes a unique subset of a climate dataset. To accomplish this, as shown in Figure 1, the VMs elect a leader at launch time; the leader maintains a synchronized queue of remote URLs (pointing to data files to analyze) from which the workers pull URLs. When a worker pulls a URL, it uses GridFTP [6] to fetch the file, runs TSTORMS analysis code on it, stages the results out to a remote directory, and pulls a new URL. Upon completion, the VMs shut themselves down.

Results: Our most encouraging preliminary result is an analysis of ~500GB of climate simulation data in under 3 hours using 34 VM instances. Analysis of the same dataset using the same code has previously taken 5-7 days running on climatologists' personal machines. We have been able to control analysis speed by using varying numbers of instances. Resolving software dependencies was simple. Some reliability issues were found with the cloud platform we used; it was often difficult to get VM instances to run. We believe that these issues will be resolved in time, and even with the issues we were able to successfully complete analysis tasks using many VM instances.

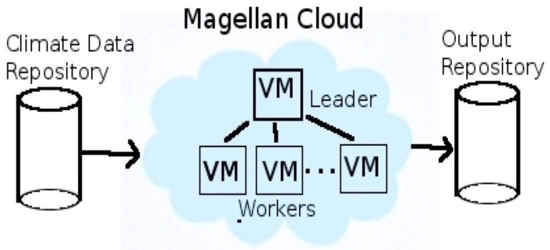

Figure 1: Architecture design of the distributed parallel virtualized climate analysis platform

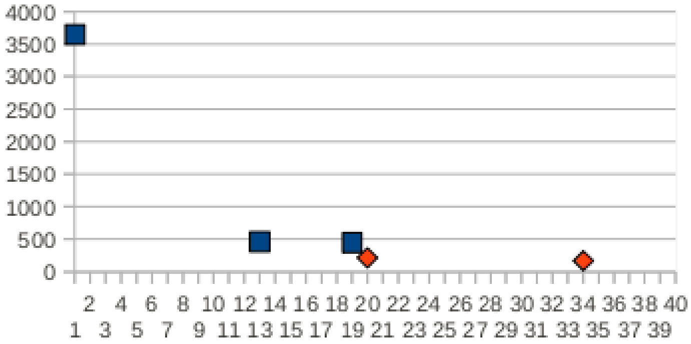

Figure 2: Total analysis time in minutes vs number of VM instances. (Blue: small resources, Red: large resources; indicates how much compute resources are available to the VM)

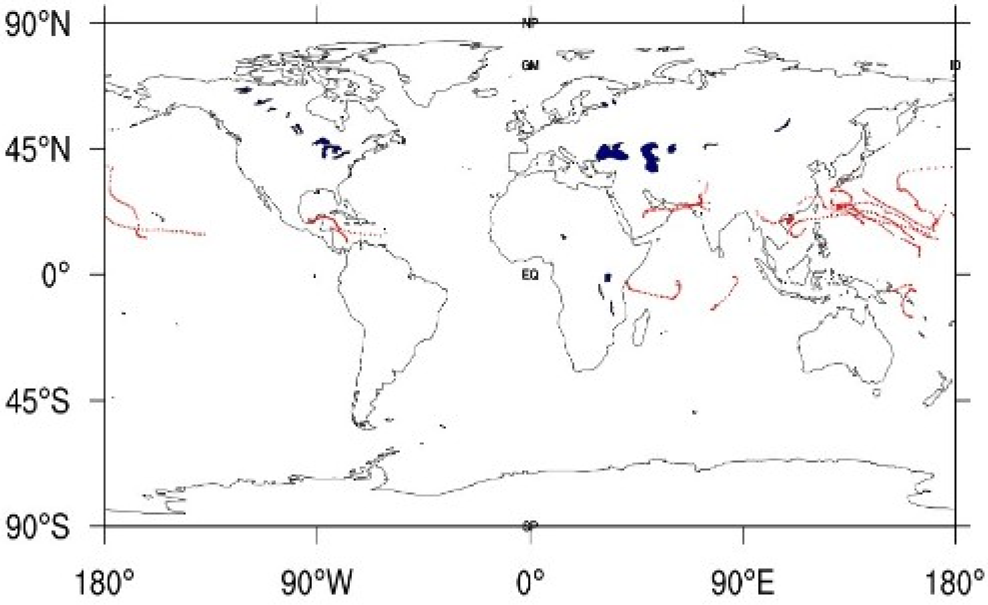

Figure 3: Plot of storm trajectories from simulated climate data analyzed on Magellan cloud by our VM. (Red dots are storms at timesteps.)

References

1. M. Cusumano. Commun. ACM 53, 4 (Apr. 2010), 27-29.
2. R. Buyya, et al. HPCC, 2008. http://dx.doi.org/10.1109/HPCC.2008.172
3. T. Knutson, et al. Bulletin of the American Meteorological Society, 2007 88:10, 1549-1565.
4. M. F. Wehner, et al. Advances in Meteorology. Volume 2010 (2010), Article ID 915303.
5. D. Nurmi, et al. In CCGRID. 124-131. 2009.
6. W. Allcock, et al. In SC05, 2005. Doi:10.1145/1105760.1105819
7. Magellan Scientific Cloud. http://magellan.alcf.anl.gov/