SAND2007-4561P

Ovis reliably monitors computers using novel parallel calculations

August, 2007

Sandia has many large computer clusters that it operates to provide answers to physics and engineering questions as part of its mission. As these clusters and the problems they solve grow in size, the probability that a critical component will suffer degradation or failure during the course of a parallel calculation run also rises. The Ovis project ironically leverages this very characteristic that increases the likelihood of failure in order to more accurately determine early warning of failure.

Capitalizing on the fact that like components should behave in a statistically similar fashion given similar environments, Ovis uses statistical and stochastic methods to characterize the distributions of values of parameters that describe the state of components in a cluster (e.g., CPU temperature, network packet loss, error rates, etc.) (see Figure 1). Values with low probability can be a sign of impending failure. This methodology allows Ovis to take into account reasonable variation in what can be considered normal (e.g., effects of manufacturing variations) and, more significantly, variations due to the environment (e.g., when CPUs near the middle of a rack run cooler than those at the top or bottom).

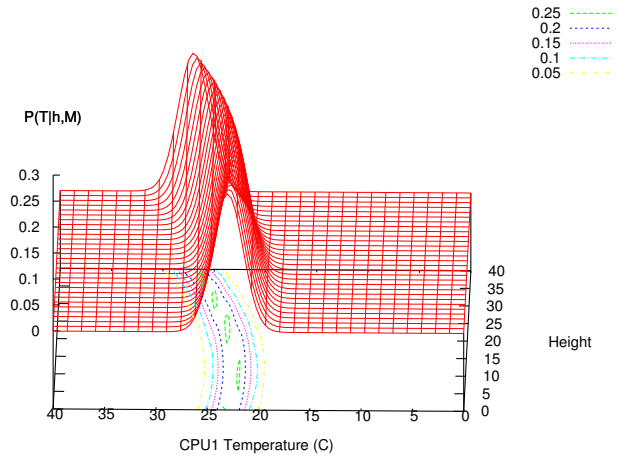

Figure 1: Ovis infers probabilistic models for system data. Here a model for normal CPU Temperature vs height takes into consideration environmental effects.

Ovis provides a set of descriptive statistics which give the system operator a mathematically rigorous

picture of the system, from which the health of the system can be gauged. It provides methods for evincing correlations between variables (e.g., the degree to which CPU temperature and CPU load are related) evaluating when diagnostics are undergoing some transition (e.g., whether a change in network traffic is significant or an expected variance), and inferring deviation from normal behavior.

Such calculations require Ovis to acquire data on many components and to perform calculations to find the best statistical model or model parameters in near real time. Additionally Ovis must present this statistical information on the cluster to people in an intuitive way (see Figure 2). Each of these tasks becomes more difficult as the cluster size increases. Ovis must itself then turn to clusters of computers for its own performance and thus ensure its own reliability as well. One key to reliability is avoidance of single points of failure, so Ovis uses a decentralized model where any computers performing the tasks above can vanish and the remainder will function properly. Within the Ovis 2 framework, this is handled by mechanisms for dynamic discovery and reconfiguration of its data collection and management resources. However, this constraint has ramifications for the way statistics are computed.

Figure 2: Ovis 2 draws information about the cluster state on a 3-D rendering to help show how the environment affects diagnostic measurements.

Many of the science applications that run on Sandia clusters can be considered *problematic* parallel computations; that is, if one single process fails to complete its part of a computation then no solution to the overall problem is provided. But one of the axioms of the Ovis project is that the problems Ovis must solve in parallel – computing descriptive, statistics, correlative statistics, or the most likely parameters for some probabilistic model – are not *problematic*. Indeed, without this axiom it would be impossible to reliably monitor large clusters. So, rather than use the Message Passing Interface (MPI) that most scientific applications use to organize parallel calculations, Ovis uses a database table to keep track of the contributions a process makes (or fails to make) to a given computation. When some processes fail to complete a given step, the results are simply computed over the samples not assigned to those processes. When only n out of N processes complete a step, the results may be higher or lower than the true result. Considering all possible sets of n processes, if some produce higher values other combinations must produce lower values; so the set of values produced by combinations of n processes must bracket the single, correct result. Determining lower limits for n that still allow Ovis to work reliably is important because these calculations will be used to change the behavior of the cluster in near real time. For example, computers likely to fail can be identified before failure actually occurs, giving science applications with *problematic* parallel calculations a chance to save their state rather than start over.

The Ovis project is investigating expanding its methodologies and framework to problems in other areas requiring anomaly detection in large data sets evinced from large numbers of similar devices such as Chem/Bio sensor grids, parallel application profilers, etc.

The Ovis (http://ovis.ca.sandia.gov) team is Philippe Pebay (8351), Jim Brandt, Ann Gentile, David

Thompson, Matthew Wong (8963) and summer intern James Jolly of Univ of Missouri-Rolla.