Toward a Better Understanding of Failures in Grid Systems

Luca Telloli∗
telloli@cs.unibo.it

Flavio Junqueira†
flavio@cs.ucsd.edu

Keith Marzullo†
marzullo@cs.ucsd.edu

Sandeep Chandra‡
chandras@sdsc.edu

Karan Bhatia‡
karan@sdsc.edu

Scientific applications, such as large-scale simulations, require a significant amount of computational resources and storage capabilities. Additionally, scientists in different locations wish to access and share data and results of computations over heterogeneous networks, such as the Internet. A solution to these demands is the interconnection of a number of cooperating, yet independent computational infrastructures into a single system, commonly known as Grid.

The large-scale nature of Grids exposes them to all kind of failures, which can impact the system in different ways: crashes can cause unavailability of a node, network problems can disconnect multiple nodes, failures of storage servers can cause processes to stall. In addition, the geographically-dispersed nature of such systems makes it harder to handle failures, as no single person has access to all the resources simultaneously. The ability to provide service despite failures is hence a crucial issue.

The first step towards this goal is to understand the nature of failures in wide-area systems. In September 2005 the San Diego Supercomputer Center and the Computer Science and Engineering department at the University of California, San Diego started a collaborative project on Failures In Grid Systems (FIGS). The project consists of collecting and analyzing failure data from the GEON grid infrastructure, and the main goals are:

• to classify common failures
• to study the dependency of failures
• to build realistic failures models
• to validate theoretical models with actual data analysis

Our methodology for collecting data is different from previous works in measurement: we collect snapshots periodically for each node and automatically parse and analyze them to extract information on failures. This allows a fine-grained classification, that includes partitions, reboots, and crashes.

We have been collecting data for eight weeks, and we expect to have sufficient material for a deeper analysis in a few more months. Figure 1 illustrates partitions and their duration during the month of February 2006, when six out of eight monitored clusters experienced some kind of failure, with over thirty partitions and a few crashes.

Figure 1: Partitions in February 2006 on GEON Grid

A preliminary analysis shows that most of the causes of unavailability are short network partitions, and that mean time to repair for crashes is significantly longer than the time for a partition to heal.

With the data that we will have in the future, we expect to confirm the trends we observed and produce failure models that capture accurately the behavior of the system, and yet are useful for the design of algorithms to achieve higher availability in systems such as GEON. These failure models will enable solutions to common problems on Grid systems, such as data replication and consistency.

∗Department of Computer Science, University of Bologna
†Department of Computer Science and Engineering, University of California, San Diego
‡San Diego Supercomputer Center