

GridS Diagnostic Scenario

1. *Header information.*

In the following scenario a group of institutions are operating such a portal for a large, geographically and administratively scattered community of Biologists (the portal in question could be doing some sort of Genome analysis, e.g. GRID-BLAST). One institution has volunteered to run the portal and also provide primary frontline support. Other sites are contributing computational and data storage resources (or portions thereof).

This community is primarily composed of faculty and their students and are spread over dozens of campuses. Users use a standard web browser as their UI to a Grid which is fronted via standard web technologies on a server. Authentication is done via password to the server, which uses the password to obtain Grid (i.e. X.509-based) credentials for the user from an online credential repository (successfully obtaining credentials implies an authorization to use the portal).

After obtaining a description of a desired task from the user, the portal software uses Grid tools to coordinate movement of data to compute resources, operate on that data by running applications on those resources, and then retrieving results to present to the user. Data is stored on resources at different sites. The computational resource used by the portal to handle a particular use request is selected at runtime from a short list (also distributed across a number of institutions contributing cycles to the portal). This selection is based on the current load on those resources.

The Grid resources at the institutions all perform authentication of the user's credential locally. Additionally some sites have their resources check with a central authorization service before proceeding (e.g. to check for revocation).

<i>User:</i>	Biology community
<i>Technical buyer:</i>	Institution IT department
<i>Economic buyer:</i>	PI at institution, funding body for Biology work.

2. *A Day in the Life (Before)*

The idea here is to describe a situation in which the user is stuck, with significant consequences for the economic buyer. The elements you need to capture are five:

- *Scene or situation:* Things are running normally the institution hosting the portal. Their help desk is monitoring local computational resources, networks, internet connection, etc. and all seem fine.

- *Moment of frustration:* At approximately 11am, the help desk begins to see a steady stream of email from users of the biology community complaining that the web portal is hanging when they attempt to submit jobs.

Staff at the help desk start pursuing various options for diagnostics including:

- Trying to use the portal themselves, but since the help desk members aren't member of the biology community, no one has an account on the portal nor credentials in the repository
 - Checking to see if the various computational and data resources that could be used by the portal are up and running, and they seem to be.
 - Checks to see if the Grid software on the resources is working, but no one on the help desk team can remember the password to their private keys since they don't use them frequently themselves.
 - Trying to verify that the network connections between the portal and the users in question are up and stable. They don't know the exact client machines the users are connecting from, but do know their institutions by inference from their email addresses, so a series of pings are started to `www.<institution name>.edu`.
 - Checking the logs on the portal. One member logs in and finds a 30MB log file. They wade through it, trying to make sense of the messages and figure out which, if any, are relevant.
- Some possible happy endings:
 - One of the help desk staff gets lucky with the pings and figures out a regional network connecting a large number of campuses is having routing flap resulting in bad connectivity between biologists on those campus and the most of the world (email gets through, but interactive stuff is hit and miss).
 - The person sorting through the log file on the portal get lucky and notices that queries to the credential repository are going unanswered. An appropriate admin is called and a bad disk is fixed.
 - A number of user's credentials were all acquired at the same time when the portal was unveiled. They expired this morning tickling a bug in the credential repository. Credentials are renewed and a bug is filed with the repository supplier. A search is done to see if other credentials will be expiring soon.
 - Someone notices that the NTP client on the credential repository has died a while back and the clock drift between the credential repository and some of the Grid resources has passed some threshold and authentication attempts to those resources are failing.

This failure is triggering a bug in the portal code which simply keeps trying thinking it's a transient error.

- A network outage has occurred between the site hosting the portal and one of the other sites providing Grid resources, resulting in traffic being black holed. That site's resources are removed from the list used by the portal.

- *Desired outcome:*
The problem is detected quickly; managers of the service fix the problem and begin a postmortem process to decide if changes in the management process of the service need to be made. The first line of support (the helpdesk) needs to have enough data to locate the problem and give the experts of the service enough data to solve the problem quickly
- *Attempted approach:* Ad hoc stabs at guessing at the problem with out access to key information or knowledge of the service. Trying to retrieve log data and check resources they didn't have access to. Bringing experts in to solve the problem.
- *Interfering factors:* Problems include unavailable data (is the user's network working?), too much data (not enough information), inability to access data and resources, unfamiliarity with technology they don't use much themselves.
- *Economic consequences:* Key employees diverted from their normal tasks to solve the problem, which costs real dollars. The trust of the biologists and funding agency is lost. Biologists may start rolling their own solutions resulting in much reinvention of wheels, different solutions and slowing the pace of their research.

3. *A Day in the Life (After)*

Now the idea is to take the same situation, and the same desired outcome, but to replay the scenario with the new technology in place. Here you just need to capture three elements:

- *New approach:*
At 8:30am the helpdesk starts getting notices from the portal service that authentications are failing to various resources. These notices themselves do not contain enough information to diagnosis the problem; just that the authentication failed for some reason which only the authenticating resources know. The failures in question span a number of different users and resources, though some resources seem over-represented. The frequency of the notices also seems to be increasing.

The help desk first tries to determine what percent of total attempts the failure represent, so they expand their visualization of notices from the portal system to include successes. The result shows that the majority of authentications are successful, but there seems to

be increasing number of failures.

The administrator of the portal system is called at this time. Calls from users of the portal start flowing in at this point reporting hangs.

In order to diagnosis the problem the help desk attempts to gain access to logs of the Grid resources at the remote sites. They have to try several before they gain access as many don't, by local policy, allow their access. They quickly scan through its error notices and find a number of failed authentications due to user credentials not be valid yet (i.e. they appear to be post-dated).

This information is reported to the system administrator of the portal, who investigates and notices that credentials from the repository are indeed valid at a point in the future that is just on the edge of the system's tolerance, resulting in sporadic failures depending on how long the portal takes between acquiring the credential and trying to use it. They connect to the repository notice clock drift due to lack of a running time daemon. They correct the clock and restart the time daemon and normal operation resumes.

At a postmortem the administrators of the portal system agree that the help desk needs access to logs related to the portal at all the remote resources that are used by the portal. Those sites are contacted and appropriate policy changes in the access control systems to those logs are made. Additional monitoring is also put into place locally for various daemons on the portal systems that are required for the health of the system.

- *Enabling factors:*
 - Standard access to event information at other sites.
 - Ability to filter such information based on severity and type (e.g. authentication).
 - Ability for help desk staff to diagnosis application they don't regularly (or ever) use themselves.
- *Economic rewards:*
 - Highly efficient operations staff that is proactive.
 - Trust from biology community and others in the institution.