Successfully Deploying a Functional caGrid Node: Technical Guidance

NCICBIIT tested the Life Sciences Distribution (LSD) bundle on the hardware and software listed in Path 1. We tested the bundle with an average of five concurrent users, using each application in the bundle. The selection of hardware and software should be driven by site-specific requirements, including size of data, number of users, and deployment architecture. *Appendix 1* provides additional guidance on sizing.

It is likely that each Center will reuse existing hardware for the initial deployment of a functional grid node. You may also consult with the caGrid Deployment Team, accessible through the Deployment Advisory Center (caBIGconnect@cancer.gov), for further guidance on deployment architectures and configuration. As Centers gather additional requirements, additional hardware resources might be required to better size the production servers.

Path 1: Install the LSD Bundle

Staffing Requirements

- An experienced tech can install the entire bundle and deploy all services on the Grid within a day. It takes an average of two to three hours to install each application in the LSD bundle.

Hardware/Software: As tested by NCICBIIT

- Hardware – (Linux: HP ProLiant DL585, AMD Opteron 852 2.4 [4 processors], 32GB memory, Red Hat Enterprise Linux 4 AS, 400 GB local storage; Windows: HP ProLiant DL585, AMD Opteron 852 2.4 [4 processors], 16GB memory, Windows 2003 Enterprise Edition Service Pack 2, 400 GB NTFS local storage)
- Software – Software and Installation documents are available on the caGrid Knowledge Center Web site at: https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Deploying_caGrid_Track.
  LSD infrastructure software:
  - Java 2 Platform Standard Edition 5.0 Update 10 (J2SE 5.0), Apache Ant, 1.7.0
  - Database (Oracle, MySQL)
  - Application server (JBoss installed automatically with each LSD application, Tomcat for MIRC)
- Installed applications – caArray (gene array data management repository), caGWAS (genome wide association repository), caTissue (biospecimen repository), NCIA (imaging tool), CTODS (clinical trials data repository)
- Ports/Security –
  - Externally accessible DNS (Internet access to your grid services)
  - 5 Ports incoming to reach services, and access to two outside URLs from the firewall. See install guide and *Appendix 2* for details.

Support available from caBIG™

- Documentation: Installation guide, End user documentation
- caGrid Deployment Team
- Webinars offered through the Deployment Advisory Center and available on caBIG™ Web site

Caution:

Ports: Opening firewall ports usually requires significant lead time (depending on IT security policies), please initiate requests to open firewall ports early in the deployment cycle.

Path 2: Install Any caBIG™ Application

Staffing Requirements

- Technical staff required for install. It takes an average of two to three hours to install an application.

Hardware/Software Guidance

- Hardware/ Software/ Installed applications – The LSD install guide (available on the caGrid Knowledge Center Web site at https://cabig-kc.nci.nih.gov/CaGrid/KC/index.php/Deploying_caGrid_Track) has links and instructions for installing each application in the LSD bundle. You can choose to install one or more applications from the LSD bundle. You will have to install the infrastructure software (Java, Ant, Database) irrespective of the application(s) you choose to install. For applications not in the LSD bundle, please refer to each application's install guide for specific guidance or contact the caGrid Deployment Team for additional guidance. Please refer to *Appendix 1* for guidance on hardware.
- Ports/Security – One per application deployed, and access to one outside URL from the firewall.

Support available from caBIG™

- Documentation: Installation guide, End user documentation
- caGrid Deployment Team
- caBIG™ Knowledge Centers
- caBIG™ Support Service Providers

Path 3: Adapt One of Your Existing Applications

Staffing Requirements

- Very technical staff. Time required will depend on the complexity of the system.
- Adaptation typically requires software modifications and, therefore, will require a full development cycle, including requirements, design and model development, caBIG™ compatibility review, development, and testing.

Hardware/Software Guidance

- Hardware/ Software/ Installed applications/ Ports/ Security – Determined by application requirements and the adaptation process

Support available from caBIG™

- caGrid User Lists
- caGrid Deployment Team
- caBIG™ Workspaces (Architecture and Vocabulary & Common Data Elements)
- caBIG™ Knowledge Centers
- caBIG™ Support Service Providers
- caBIG™ program staff will work with you individually to support adaptation of your chosen application

NOTES

1. For some guidance on sizing computing resources for your site, see *Appendix 1* below.
2. For the list of ports and examples of configurations see *Appendix 2* below.

Appendix 1: Guidelines for Sizing Hardware

caBIG™ applications are n-tier applications and can be configured on multiple servers to take advantage of a scalable architecture.

A typical installation has two servers:

- An application server
- A database server

An existing database can be used for most applications, and a new database install is not required. The application server must have access to the Internet with the required ports open for incoming and outgoing traffic.

The LSD bundle is architected as a set of independent applications with the capability to share a single database and single User Provisioning Tool (UPT). Each application runs in its own container (application server). caArray and caTissue use a different container for the grid services. The shared UPT runs in a separate container.

Microarray and Imaging data are usually large data sets (possibly Gigabytes) and require more memory. In 32-bit machine, the addressable memory is 2 GB for a container. Therefore, applications are deployed in different containers to avoid memory contention.

The following table shows the default memory allocation for applications in the LSD bundle:

Component	Version	JVM COL	Heap Size	JVM SizE	Total Memory (MB)
Operating Syst					1,024
caTissue	1.2.2	2	1024	256	2,560
caArray	2.0.2	1	2048	256	2,304
caGWAS	1.0.0	1	2048	256	2,304
CTODS	1.0.0	1	512	256	768
NCIA	3.0.0	1	512	256	768
UPT	3.2.0	1	512	256	768
					10,496

Caution: If the database resides on the same machine, additional memory needs to be allocated for the database.

Processors:

The required processing power is largely determined by the number of users and size of data set. For low usage and activity, a single processor machine might be sufficient. For an active user population manipulating large data sets, multiple processors or machines may be required. However, upgrades can be done after an initial installation and functional grid node is established, at the Center's discretion.

Based on the preceding information, the IT support staff should be able to select a reasonable machine to support the Center's initial deployment needs.

Appendix 2: caGrid and caBIG™ Tools Deployment Preparation

Requirements for Internet Access

Introduction

When a research institution adopts caBIG™ tools such as the LSD bundle or one of its components for participation in the national caBIG™ research grid, the grid services must be accessible to systems from other institutions on the grid. To be available on the national grid, the service must be reachable from outside the firewall that typically protects the institution's internal network and users. This external access is achieved by configuring the security on the computer and network hosting the caBIG™ local grid node or application in such a way that one or more ports are open and the remainder of the network is not vulnerable to external accesses.

Opening a port for external access typically requires an Institution Security Officer review and justification documentation, followed by modifications to the firewall rules to open the ports. This process may take time and until completed, the grid service will not be available on the national grid.

This paper is intended to facilitate communication between the caBIG™ Deployment Lead and the Security Officer to prepare for the installation of the caBIG™ grid node and tools. Complete installation requirements and installation support is available through caBIG™ documentation, Knowledge Centers and the caGrid Deployment Team.

Scope

This document provides information for the manager of Network Security with firewall rule changes for a caBIG™ application installation at the research institution deploying the tool. The default configurations and default ports required for the installation are listed. The specific port numbers are configurable items, as indicated, at the time of installation. There are many configurations possible, based on the requirements of the institution. The security officer and network managers may chose a specific configuration. This paper lists the ports needed for access, based on sample topologies.

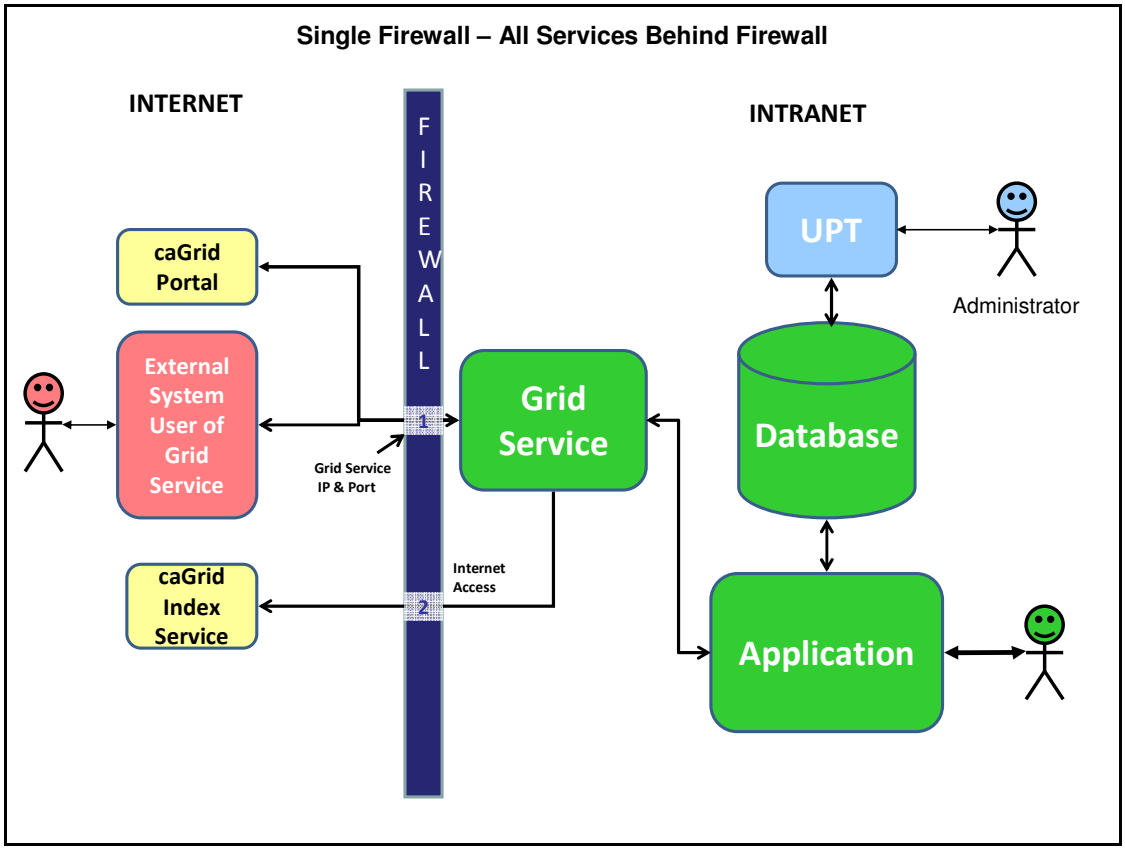
Description

This section provides an overview of typical deployments of an application. The diagrams are provided as illustrations for the firewall rule changes regarding the ports and DNS names that need to be made. For most grid services, at least 2 ports need to be opened: 1) The port for external users to access the local grid service; and 2) the outgoing port for Internet access for the local service to update national caBIG™ grid Index.

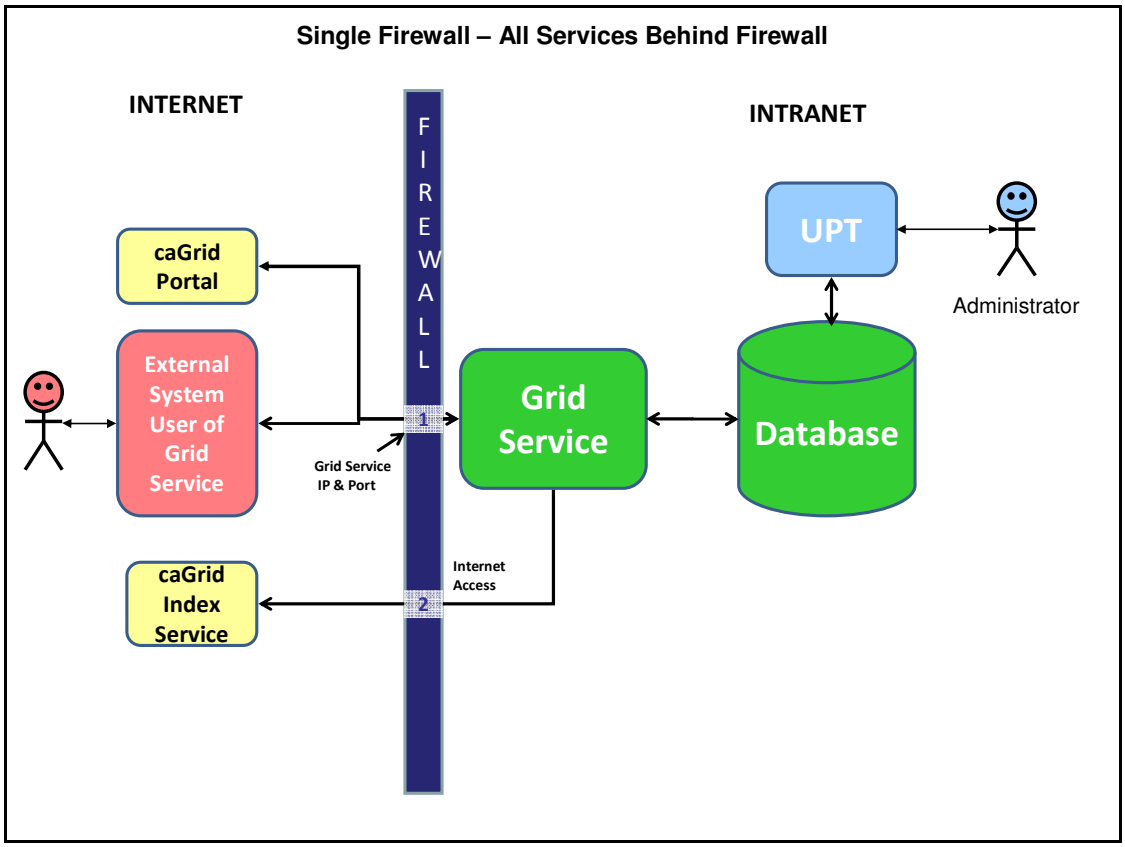
The caBIG™ applications are built as N-tier applications and can be physically deployed in multiple tiers/machines. A grid service may access a database directly or may use the application logic to perform the service. A grid service may not have a companion application. In the following diagrams, only one example of a configuration is given of a grid service accessing the database directly. The specific configuration depends on each application and grid service, and the exact configurations will be documented in the installation guides of each tool.

Single Firewall

In this configuration, all services are behind the single firewall. Only the mandatory ports need to be configured.

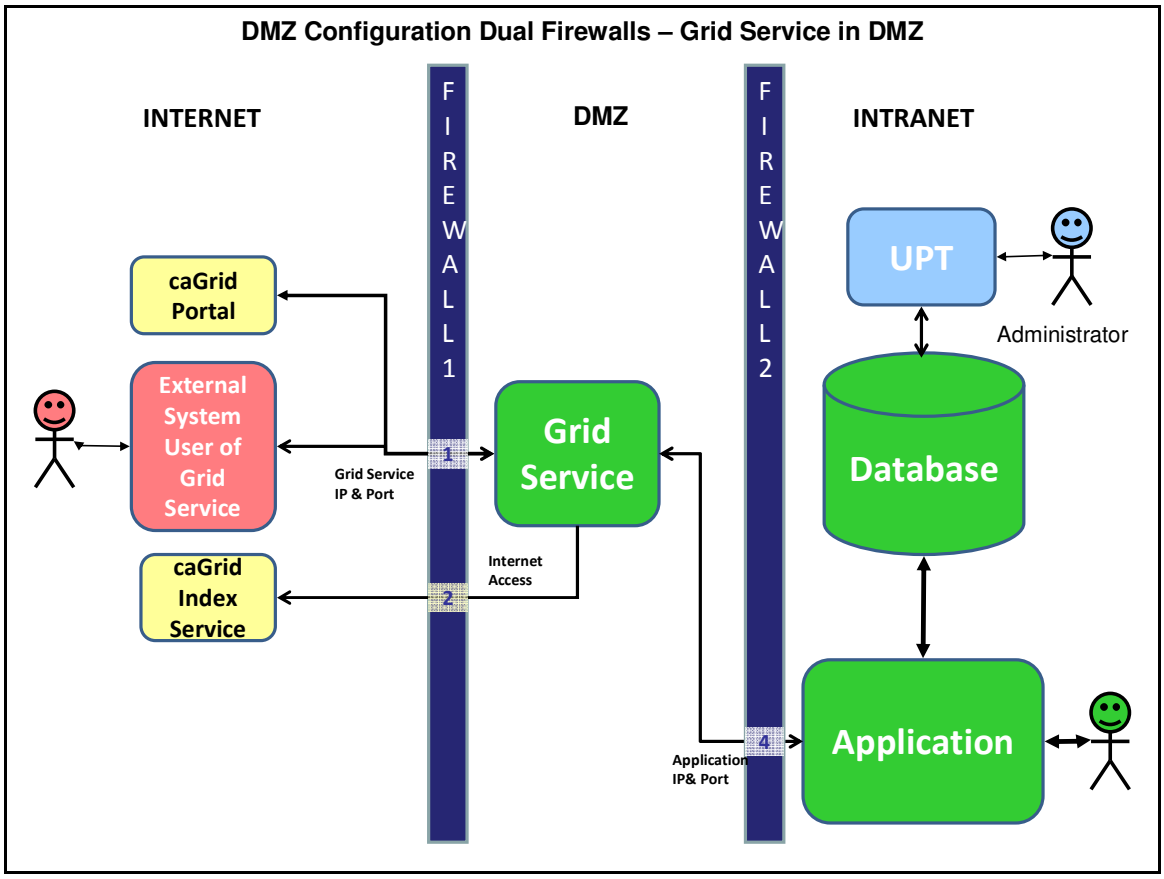


The following configuration shows a grid service accessing the database directly.



Dual Firewall

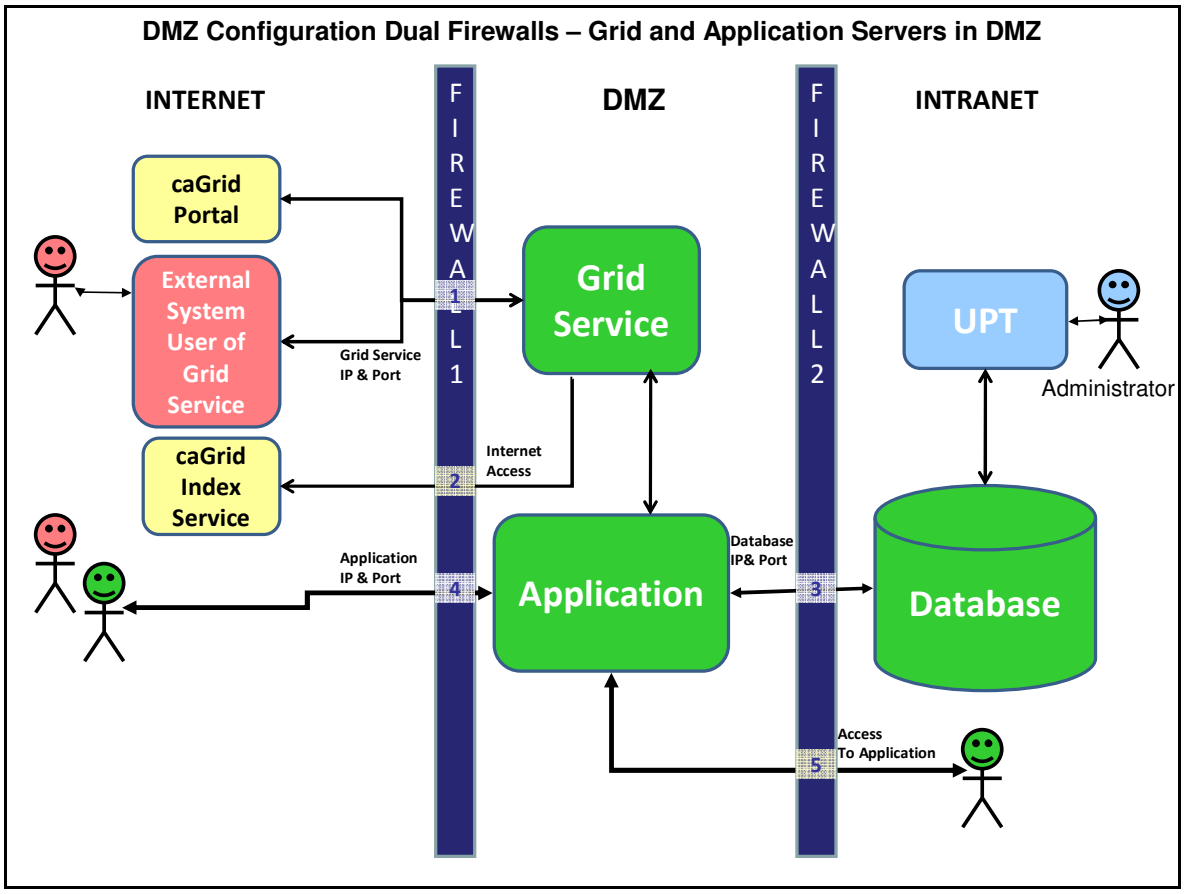
The use of two firewalls creates a "DMZ" and further restricts access to the institution's internal systems. Typically the database is behind the second firewall and the grid service/ application will need access to the database port based on the two scenarios shown below:



Note: If the grid service accesses the database directly, then the database port also would need to be opened in this configuration. In this configuration, only internal users can access the application, while external users can access the grid service.

Dual Firewall with External Access to the Application

When an application is shared with users outside the institution, then the application port also needs to be made available through the firewall. If the application server is placed in the DMZ, then internal users need access either via general Internet access or via firewall rules for internal users.



In this configuration, the database port needs to be opened. Both internal and external users can access the application.

Summary of access ports represented in the diagrams

Number in diagram	Description
1	Port for external systems to access the local caGrid Service
2	Port for the local caGrid service to access the Internet and the national caGrid Index service
3	Port to access the database
4,5	Ports to access the user application component

The User Provisioning Tool (UPT)

caArray and other caBIG™ tools use the User Provisioning Tool (UPT) to manage user accounts and access authorizations. This tool is an application that may be placed inside or outside the perimeter firewall. If the tool is installed on a server inside the firewall, only administrators located inside the firewall (or using a VPN access through the firewall) may use the UPT. In a collaborative situation where some administrators may be located outside the firewall (for example in another institution), then the UPT tool may be made accessible by opening the following optional port: http://<IP_or_DNS>:**46210**/upt

List of Ports for caBIG™ Grid Nodes and Selected Tools

The following table summarizes services and ports for selected caBIG™ applications.

NOTE: The local grid services need to access the caGrid Index service as an outgoing request. Multiple grid services, one from each application, access the same Index service. The firewall needs to be configured to allow this outgoing request. See URLs below.

Component/ Service	Description	Configurable Mand/Opt # in diagram	Default Number	URL
caArray				
caArray grid service	Primary port for incoming access to the locally hosted caArray grid service Note: The caGrid Portal will access the same service and port	Configurable Mandatory #1	18080	http(s)://<IP_or_DNS>:18080 /wsrf/services/cagrid/CaArraySvc
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Web user of the UPT tool	Administrator access to user authorization management tool	Configurable Optional	46210	http://<IP_or_DNS>:46210/upt
Web user of the caArray application	Port needed by the caArray application graphical user interface component	Configurable Optional #4,5	38080	http://<IP_or_DNS>:38080/caarray
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521
caTissue				
caTissue service	Primary port for incoming access to the locally hosted caTissue service Note: The caGrid Portal will access the same service and port	Configurable Mandatory #1	47210	http(s)://<IP_or_DNS>:47210/wsrf/services/cagrid/CaTissueCore
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Web user of the caTissue application	Port needed by the caTissue application graphical user interface	Configurable Optional #4,5	43210	http://<IP_or_DNS>:43210/catissuecore

Component/ Service	Description	Configurable Mand/Opt # in diagram	Default Number	URL
	component			
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521
UPT				
User Provisioning Tool	A port to allow access to the user management tool from outside the firewalls	Configurable Optional	46210	http://<IP_or_DNS>:46210/upt
NCIA				
NCIA grid service	Primary port for incoming access to the locally hosted NCIA grid service Note: The caGrid Portal will access the same service and port	Configurable Mandatory #1	45210	http(s)://<IP or DNS>:45210/wsrf/services/cagrid/NciaCoreService
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Image Server	The local MIRC application should be able to access the image server hosted by NCI	Configurable Mandatory #2	443	Internet access from the local application required https://imaging.nci.nih.gov
Web user of the UPT tool	Administrator access to user authorization management tool	Configurable Optional	46210	http://<IP_or_DNS>:46210/upt
Web user of the NCIA application	Port needed by the NCIA application graphical user interface component	Configurable Optional #4,5	45210	http://<IP or DNS>:45210/ncia
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521
MIRC for NCIA	Port needed by the MIRC application graphical user interface component	Configurable Optional #4,5	58080	If Oracle is used as the underlying DB
			58081	If MySQL is used as the underlying DB
CTODS				
CTODS grid service	Primary port for incoming access to the locally hosted CTODS grid service	Configurable Mandatory #1	44210	http(s)://<IP or DNS>:44210/wsrf/services/cagrid/Ctods

Component/ Service	Description	Configurable Mand/Opt # in diagram	Default Number	URL
	Note: The caGrid Portal will access the same service and port			
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Web user of the UPT tool	Administrator access to user authorization management tool	Configurable Optional	46210	http://<IP_or_DNS>:46210/upt
Web user of the CTODS Viewer application	Port needed by the CTODS Viewer application graphical user interface component	Configurable Optional #4,5	44210	http://<IP or DNS>:44210/CTODSViewer
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521
caGWAS				
caGWAS grid service	Primary port for incoming access to the locally hosted caGWAS grid service Note: The caGrid Portal will access the same service and port	Configurable Mandatory #1	42210	http(s)://<IP or DNS>:42210/wsrf/services/cagrid/CAGWAS
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Web user of the UPT tool	Administrator access to user authorization management tool	Configurable Optional	46210	http://<IP_or_DNS>:46210/upt
Web user of the caGWAS application	Port needed by the caGWAS application graphical user interface component	Configurable Optional #4,5	42210	http://<IP or DNS>:42210/cagwas
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521
caGrid node				
caGrid service	Primary port for incoming access to the locally hosted grid service Note: The caGrid Portal	Configurable Mandatory #1	non Secured 8080	http(s)://<IP_or_DNS>:**<PORT>**/wsrf/services/cagrid/<SERVICE_NAME> NON-SECURED Tomcat, Globus, or any other server: **8080**

Component/ Service	Description	Configurable Mand/Opt # in diagram	Default Number	URL
	will access the same service and port		Secured 8443	SECURED Tomcat, Globus, or any other server: **8443**
caGrid Index Service	The local caGrid Service updates the caGrid Index at regular intervals to indicate that the service is available and active	Configurable Mandatory #2	Specific URL Access	The Machine hosting the local service should be able to access http://cagrid-index.nci.nih.gov:8080/wsrf/services/DefaultIndexService
Database access	The port is used to connect to the database supporting the application or grid node	Configurable Optional #3	3306	The default port for MySQL is 3306
			1521	The default port for Oracle is 1521