CHAPTER-1

Introduction

This chapter provides an introduction to a new paradigm called Data Intensive Science. Concepts of remote sensing and earth observations systems are discussed. Survey is carried out on software architectures used for satellite image processing and requirements of future earth observation systems. Technology trends being adopted by major space agencies to cater to future earth observation requirements are reviewed. Motivations for carrying out this research are discussed in detail. This chapter concludes by providing the outline of research contents.
CHAPTER ONE: INTRODUCTION

Remote Sensing Satellites, Large Hadron Collider (LHC), Computer Generated Animations, Protein Folding Experiments, Climate Modeling and Medical Imaging generate huge volume of scientific data. Potential of this large volume of scientific data to initiate new research has evolved a new paradigm called Data Intensive Science (Szalay 2011). This paradigm has considerably changed the ways in which research and innovations are being carried out in the field of data processing and analysis.

Remote sensing is the acquisition of information about an object or phenomenon, without making physical contact with the object. Remote sensing satellites are used for observing earth and other planetary bodies using different regions of electromagnetic spectrum such as optical and microwave. Earth observation systems consists of constellation of remote sensing satellites and necessary ground based infrastructure to receive, process and disseminate remote sensing data. Earth observation data processing is both compute and I/O intensive as compared to conventional data processing, which is either compute or I/O intensive.

Earth observation systems have continuous rapid growth and hence our research is focused on earth observation user centric optical data processing. Developments in sensors, electronics and optics have made it possible to acquire sub meter spatial resolution data to cater to cartographic requirements. Hyper-spectral data with large number of narrow channels support applications such as mineralogy and crop identification. Data with large coverage (swath) and improved revisit time support
requirements of natural resource monitoring. Earth observation data is becoming increasingly voluminous and hence unmovable.

In current scenario, earth observation data are downloaded at user end and then user specific processing and analysis is carried out. Requirement of computational resource, specialized software for processing and large network bandwidth for downloading are limiting factors in utilization of earth observation data. Traditional software architecture for earth observation data processing is based on monolithic and object-oriented architecture (Booch, et al. 2007). Software is developed based on requirements, either by using existing components or by developing software from scratch using edit-compile-link model (Turner, Budgen and Brereton 2003). This requires specialized professionals for software development.

Most of the end users of earth observation data are application scientist, has a limited knowledge of software. It is difficult for such users to find out and use specialized software components available with the earth observation data providers. Analysis and processing of data at user-end requires access to computational resources. Nominal computational loads, such as processing of a single scene, can be taken care by local resources at user end. Users find it difficult to handle bulk data and meet the processing requirements of computational intensive applications.

Earth Observation applications can broadly be classified under following categories.

a) Data centric applications: This type of applications requires access to huge volume of data extending to months and years. Time series data processing and Temporal Binning of satellite data falls under this category. Satellite is able to acquire data with
systematic coverage during its mission life and continuity in mission ensures systematic data availability for long time. This has made satellite data very popular among researchers for changes management related studies and for deriving statistical trends for a given area and required time period.

b) Compute centric applications: These user applications require a huge computational power. Image ortho-rectification, data fusion from multiple satellites; rule based classification and hierarchical image matching can be classified in this category. Applications requiring reprocessing of satellite data with improved algorithms and techniques also generate extra computational loads.

c) Near real time applications: User interested in monitoring a specific region requires data to be processed with minimum latency. The value of the derived information for such applications decreases with time. Applications related to tracking of events such as cyclone, forest fires, oil spill falls under this category.

Applications such as time series analysis requires processing of sequence of images acquired over a common area at successive time interval. This requires additional computational resources. Availability of computational resources for handling application requirements becomes bottleneck for processing earth observation data at user end. As the current architecture is not adequate to meet the user centric data processing requirements and hence there is a need of studying and exploring possibility of adopting new architecture for processing earth observation data.

Potential areas of research related to earth observation data processing includes data processing architecture, process scheduling techniques, discovery of data and services, dissemination of data and dynamic application generation using workflow
composition. As part of this research work, we propose a software framework for user-centric data processing, with a goal of using grid services extensively for applications related to earth observation data. This is a new paradigm with respect to the conventional data processing architecture.

**Grid Services for Earth Observation Image Data Processing (GEO-ID),** developed as part of this research allows users to create on-demand applications and execute them on the virtual environment using specialized services related to earth observation data processing. GEO-ID provides users, access to earth observation data and collocated computational resources. This helps in processing data with quick turn-around time.

**1.1 SATELLITE DATA PROCESSING OVERVIEW**

Data Processing System consist of hardware and software for acquisition and processing of data received from the remote sensing satellites. Data Processing consists of two major steps called pre and post processing. Data Products Generation Software (DPGS) takes care of pre processing and Value Added Data Products Software (VADS) carry out post processing. DPGS is responsible for performing standard level of correction on the satellite data in terms of radiometric and geometric corrections. VADS carries out extra processing such as accuracy improvement, mosaic, merge, ortho-rectification and other operations based on application specific requirements. These processed data are called as data product. Overview of the Satellite Image Processing is shown in Figure 1-1.
Satellite data is acquired by the data acquisition system, which archives the raw data at a central data archive. Data acquisition system also generates browse and accession catalogue. Browse is a sub-sampled product chip and accession catalogue is the product metadata. This information is used by the Information Management System and helps the end users to order products. Based on user request pre-processing of the raw data is carried out at DPGS. The processed data is converted into products and are stored at product archive. These products undergo a data quality check and are delivered to end users. Output of DPGS based on user requirement is also transferred to VADS for post processing. Data quality evaluation is carried out on routine basis to ascertain that products comply with the mission specifications.

Software architecture used for processing satellite data as shown in Figure 1-2 uses edit-compile-link model (Turner, Budgen and Brereton 2003) for software
development. The end results of data processing are data products, which are supplied to end users.

![Data Processing Architecture](image)

**Figure 1-2 Data Processing Architecture**

### 1.1.1 Data Pre-Processing

The satellite data pre-processing consist of radiometric and geometric corrections applied on the raw satellite data. Radiometric correction of remotely sensed data involves the processing of digital images to improve the fidelity of the brightness value magnitudes as opposed to geometric correction, which involves improving the fidelity of relative spatial or absolute location aspects of image brightness values.

a) Radiometric Corrections

The main purpose for applying radiometric corrections is to reduce the influence of errors or inconsistencies in image brightness values that may limit one's ability to interpret or quantitatively process and analyze digital remotely sensed images. Sensor on board spacecraft observes the emitted or reflected electro-magnetic energy, which does not coincide with the energy emitted or reflected from the same object observed from a short distance. Radiometric distortions are included in the data due to various elements of
the data acquisition chain as shown in Figure 1-3. Table 1-1 lists out corrections, applied as part of radiometric corrections.

![Figure 1-3 Electromagnetic Energy Travel Path](image)

### Table 1-1 Type of Radiometric Corrections

<table>
<thead>
<tr>
<th>Sub-System</th>
<th>Type of Radiometric correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCD response variability</td>
<td>Photo Response Non Uniformity (PRNU)</td>
</tr>
<tr>
<td>Optics</td>
<td>Point Spread Function (PSF)</td>
</tr>
<tr>
<td>Platform Motion and Dynamics</td>
<td>Point Spread Function (PSF)</td>
</tr>
<tr>
<td>CCD Charge</td>
<td>Photon Noise</td>
</tr>
<tr>
<td>A/D conversion</td>
<td>Quantization Noise</td>
</tr>
<tr>
<td>Placement of CCD</td>
<td>Stagger in Image</td>
</tr>
<tr>
<td>Compression</td>
<td>Artefacts in Images</td>
</tr>
<tr>
<td>Downlink</td>
<td>Pixel Dropouts or Line Loss</td>
</tr>
</tbody>
</table>

b) Geometric Correction

Geometric correction is needed to determine the correct ground position of a point visible in an image acquired from the sensor on-board the satellite. Geometric corrections compensate for distortions introduced by a variety of factors and correct imagery so as to have the geometric integrity of map. Let us call the corrected image space as output space. Geometric correction establishes a mapping between the input space (image that is acquired and radiometrically corrected) and the output space. Geometric correction comprises of two essential steps that are spatial mapping and resampling.

Spatial mapping can be achieved with different methods ranging from the true mathematical description of imaging geometry to simple polynomial mapping. There are
two categories of distortions to be modeled to establish the mapping between the input space and the output space. These are systematic and random distortions. The systematic distortions are predictable and can be modeled with true mathematical description of imaging. The random distortions are corrected statistically by comparing the modeled output with Ground Control Points (GCP). The output space is translated, rotated and scaled version of the input image. The translation, rotation and scaling are not same for the whole image but varies from pixel to pixel in an image.

1.1.2 Data Post-Processing

Data products need to meet stringent geometric accuracy specifications of intended user applications. Geo-referencing is the basic processing step towards achieving this goal. Having known the imaging geometry, the mathematical models built with the use of orbit and attitude information of the spacecraft can correct the remote sensing data for its geometric degradations only up to system level accuracy. The uncertainties in the orbit and attitude information will not allow the geometric correction model to generate products of high accuracy that can meet user requirements. Ground Control Points (GCP) are used as reference geo-location landmarks and a Digital Elevation Model (DEM) is used for correcting errors due to terrain. Such products are called as ortho-rectified value added products.

Long term analysis (time series analysis) using earth observation data requires all images to be registered to a standard template. This could be a user supplied image or can be the first image of the stack. Such products are called as template registered products. Current user requirements also call for generating products using data from different sensors (Data Fusion) and generating products with different dates and sensors over large
area (Large Area Mosaics). These products are value added data products and the associated corrections are part of data post processing.

1.1.3 Levels of Data Products

Data Products are classified into following levels based on type of corrections applied on the end products.

- **LEVEL-1 Products:** These are basic data products, which have three sub levels as shown in Figure 1-4; L1A is the RAW products and contains counts as received by the satellite along with the ancillary information. L1B are the radiometrically corrected products and L1C are radiometrically and geometrically corrected products.

![Figure 1-4 Product Levels L1A (First), L1B (Second), L1C (Third)](image)

- **LEVEL-2 Products:** These are Geo-physical parameters as shown in Figure 1-5, these products are derived from the basic L1B products; L2B is geometrically calibrated geo-physical parameter i.e. each pixel has associated geographic latitude and longitude information, which can be used for carrying out correction at user end. L2C is geometrically corrected Geophysical-Parameter.
LEVEL-3 Products: These are Binned products as shown in Figure 1-6. Binning schemes can be classified into spatial and temporal binning. In spatial binning data from multiple pixel locations are combined together to generate a coarse level of data. Temporal binning combines data acquired over different time. Examples of temporal binning include Weekly binned, Monthly binned and Yearly binned products.

LEVEL-4 Products: These are special products, which contains model output
1.1.4 Data Product Formats

Processed data is provided to users in a specific format. Some of the most commonly used format by satellite data providers includes, GeoTIFF, Hierarchical Data Format (HDF) and netCDF. These formats are supported by almost all image visualization, analysis and post processing software packages.

a) HDF data can be provided in two data management formats (HDF4 and HDF5). HDF provides a set of libraries, a modular data browser/editor, associated tools and utilities. Both HDF4 and HDF5 are general scientific formats, which can be adopted for virtually any scientific or engineering application.

b) GeoTIFF represents an effort by over 160 different remote sensing, GIS, cartographic, and surveying related companies and organizations to establish a TIFF based interchange format for geo-referenced raster imagery.

c) NetCDF (Network Common Data Form) is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data.

1.1.5 Data Product Volumes and Processing Loads

Data Product volume is dependent on the sensor and mission specifications. The processing load varies with type of corrections applied on raw data. Indian earth observation satellites acquires about 90GB of data (Level-0) per ground station per day, these include high resolution data from Cartosat series, Hyper spectral data from Indian Mini Satellites (IMS), Ocean Color Monitors (OCM) data from Oceansat-2 and medium resolution data from Resourcesat-2 satellite. This data when processed generates about 413GB (Level-1) and 80GB (Level-2) products. Processing load of 1.4 GFLOPS is
required for converting Level-0 to Level-1 products, major steps involved in this correction includes application of radiometric lookup tables, geometric correction and resampling of data. An approximate computational load of 3.2 GFLOPS is required for converting Level-1 to Level-2 products; this includes execution of product specific algorithms for generation of geo-physical products.

1.2 LIMITATIONS OF CURRENT DATA PROCESSING ARCHITECTURE

1.2.1 Processing Data based on User Specific Requirements

Data Processing software carries out pre-processing of the raw satellite data with specific options and pre-defined processing parameters, based on user requirements and mission specifications. User requires processing of data with options specific to application requirements, which in current scenario can only be done by downloading and processing the raw data at user end.

1.2.2 Capability for Discovery

Current software architecture does not provide capability to discover software elements for processing the satellite data. A user who needs to use the existing capabilities should be aware of software source and their associated interfaces. This is a limiting factor for general data users as most of them are not software programmers but still would like to process the data in a specific way by using the existing software components.

1.2.3 Resource Utilization

Data Processing Systems are designed based on the nominal load; there are situations, which require processing of bulk data. Current software is system centric and does not have a mechanism for utilizing organization based resources.
1.2.4 *Data Dissemination*

The current mechanism of dissemination provides data to end user on media or network on request/demand. This has inherent limitations as manual interaction is required to feed data to end user applications. This introduces time lag for analysis and processing of data. As applications require data in a specific format and hence current applications are format dependent. End users have to know the product format for extracting area of his interest for carrying out processing.

1.3 *TECHNOLOGY ROADMAP*

In order to support earth observation user requirements, there is a need to study technology trends and roadmap. It is required to re-look on software architectures and identify associated research areas, which can help in evolving data processing architecture to support user requirements.

The current and projected future requirements for 5th generation of earth observation systems as shown in Table 1-2 indicates that there is a requirement of a software framework, where user can generate dynamic applications based on processing requirements and run them on computational resources co-located with the data. These capabilities can be achieved by adopting grid computing and Service Oriented Architecture (SOA). Grid computing provides high throughput computing environment and better resource utilization.
Table 1-2 Earth Observation System Characteristics

<table>
<thead>
<tr>
<th>Earth Observation Generations</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Characteristics 8W Photographs</td>
<td>B&amp;W</td>
<td>30meter</td>
<td>10meter</td>
<td>1-3 meters</td>
<td>Sub-meter</td>
</tr>
<tr>
<td></td>
<td>Multi-spectral</td>
<td>with Stereo</td>
<td>Hyper spectral</td>
<td>Agile Viewing</td>
<td></td>
</tr>
<tr>
<td>Software Architecture</td>
<td>Visual Interpretation</td>
<td>Modular Programming</td>
<td>Object Oriented</td>
<td>Component Based Architecture</td>
<td>Service Oriented Architecture</td>
</tr>
<tr>
<td>User Requirements</td>
<td>Single Scene Processing with Standard Level of Products</td>
<td>Multiple Scenes with Standard Level of Products</td>
<td>Multiple Scenes with Improved Accuracy</td>
<td>Global Area Processing with Ortho rectified Images</td>
<td>Time Series Processing with Climate Quality</td>
</tr>
<tr>
<td></td>
<td>Processing with Standard Parameters</td>
<td>Processing with User Specific Parameters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Access</td>
<td>Archived data on media with latency</td>
<td>Real time data on network</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cloud computing (Vouk 2008) is one of the contemporary fast evolving technologies, which provide access to centralized resources. These resources can be grouped into three categories, i.e. Platform as a Service (PaaS), Software as a Service (SaaS) and Infrastructure as a Service (IaaS). Cloud computing provides resources from a central pool and hence requires the earth observation data to be transferred to a central location for processing. Earth observation data acquisition is inherently distributed, data volumes are exponentially growing and demand for applications dealing with global data are increasing and hence it becomes practically impossible to transfer data to a cloud.
In subsequent sections we discuss service oriented architecture and grid computing technologies and carry out literature survey of research related to adopting these technologies for earth observation data processing and analysis.

**1.3.1 Service Oriented Architecture**

1.3.1.1 Definition

Service Oriented Architecture (SOA) is defined as a “set of components which can be invoked, and whose interface descriptions can be published and discovered”

SOA is also defined as “an architectural discipline that centers on the notion that IT assets are described and exposed as services. These services can be composed in a loosely coupled fashion into higher-level business processes, which provide business agility and capability to address issues of IT heterogeneity”

SOA in business terms is defined as “A design methodology aimed at maximizing the reuse of application-neutral services to increase IT adaptability and efficiency”

1.3.1.2 Generic Model for Service Oriented Architecture

A client, service provider, and a service broker constitute a generic model for Service Oriented Architecture (SOA) as shown in Figure 1-7. A service provider willing to provide functionality in the form of service needs to registers with a service broker by providing a set of parameters, which describes the service. The end user client, who is interested in utilizing the functionality, will first check up with the service broker regarding availability of service. Once the service is discovered, the client based on information provided by the service broker can use the service.
1.3.1.3 Functional Elements of Service Oriented Architecture

**Discovery:** Service Oriented Architecture provides capability to locate the required service based on user requirements. Service broker helps in locating the service. The method of discovery of service can be built in the service-oriented architecture using repositories such as UDDI, which allows the end user client to search service based on keys such as service name, provider organization and taxonomy classifications. Other discovery standard such as ebXML concentrates on definition of business abstractions and its specifications. DAML-S provides enhanced options, where service can be described under three heads, Service Profile, which describes what service does, Service Model which describes how service works and Service Grounding, which describes how to use the service.

**Description:** Service Oriented Architecture provides capability for the service to describe its functionality using a standard interface called Web Service Description Language (WSDL); this interface helps in advertising aspects related to invocation of service.

**Delivery:** Service Oriented Architecture provides a mechanism for invocation of service. The end user after having discovered the service can use the service description for invocation of service. Service invocation involves, specifying input and output
parameters and mechanism for service termination. Service can be invoked with static or dynamic interface. Static invocation requires one time compilation with the stubs, while in case of dynamic invocation; the necessary information required for invocation of the service is built at run-time.

1.3.1.4 Implementation Scenarios

Service Oriented Architecture can be implemented by using Web Services; Broker based techniques such as CORBA, Network Centric approaches such as JINI or by using Grid Services. The possible implementation scenarios for Service Oriented Architecture are shown in Figure 1-8.

![Figure 1-8 Implementation Scenarios of Service Oriented Architecture](image)

Broker based architecture was first used to build Service Oriented Architecture but could not succeed due to issues of interoperability and tight-coupled integration. Network Centric approach JINI has a limitation that only JAVA based applications can use these services. Web services are one of the most common and popular way to realize Service Oriented Architecture. Web services as are built using popular internet protocols, are highly interoperable, and provide a loose coupling integration. Open Grid Service Architecture (OGSA) aims to define a common standard and architecture for grid-based applications. OGSA uses Web Services Resource Framework (WSRF) for providing stateful web services and comply with WS-specification as shown in Figure 1-9.
Web Service Specifications can broadly be classified into categories as shown in Figure 1-10. Messaging, Security, Transaction and Metadata are four categories of web service specifications. Messaging includes WS-Addressing, which specifies how to address resources in a web services environment. WS-Transfer provides specifications regarding data transfer. The Security Specifications include WS-Security, WS-Trust and WS-Federation. Transaction specifications include WS-Coordination and WS-Atomic Transaction. Metadata Specification includes WS-Policy and WS-Discovery.

1.3.1.5 Layered View of Service Oriented Architecture

Service Oriented Architecture (SOA) follows a layered architecture as shown in Figure 1-11. The Network layer, which is the underlying transport protocol layer, is used
for establishing the connectivity fabric. One of the most popular protocols used for transfer of data over Network layer is Hyper Text Transfer Protocol (HTTP). The Messaging Layer isolates the end user applications from the Network layer and helps in describing the service semantics. Extended Mark-up Language (XML) is the messaging formats used for communication of documents and procedure calls. Simple Object Access Protocol (SOAP) is XML based messaging protocol, which is used widely in web services.

Table 1-3 compares strength and weaknesses of SOA.

![Layered View of Service Oriented Architecture](image)

### Table 1-3 SOA Strengths and Weaknesses

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weaknesses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Provides capability for dynamic discovery of software and can be used across different platforms and architectures</td>
<td>Overheads of encoding and decoding for data transfer and does not scale well for large scale distributed resources</td>
</tr>
</tbody>
</table>
1.3.2 Grid Computing Architecture

Grid is the computing and resource management infrastructure, which allows user to view a set of large-scale, distributed resources as a virtual organization and creates an environment, where user can access these resources seamlessly (Foster, Kesselman and Tuecke 2001). Building block of Grid includes, network, computational resources and data archival resources. Network is used to link the geographically distributed resources, computation resources are used for processing and data archival resources are used to archive and share huge volume of distributed data. Grid can be classified as computational, data and application grid.

The emphasis of computational grid is on discovery and scheduling of processes on configurations, which improves the throughput of an application. Information Power Grid (IPG) developed by NASA is an attempt for building high performance computational grid. DataGrid is a project funded by European Union, with an aim to build next generation infrastructure for computation and analysis of large-scale databases. DataGrid has identified earth observation applications as one of the potential application for the project.

In order to standardize the development of grid based applications the Open Grid Service Architecture (OGSA) is developed by Globus Grid Forum (GGF), which aims to define a common standard and open architecture for grid based applications. OGSA defines the basic behaviour of the service but does not put any requirements on how or what should be performed in the services. OGSA requires that the services should be stateful. Web Service Resource Framework (WSRF) defined by OASIS provides
standard for making services stateful. Globus Toolkit (Foster and Kesselman 1997) is one of the most popular toolkits for implementing grid. Relationship between these components is shown in Figure 1-12.

![Diagram](image.png)

**Figure 1-12 Relationship between OGSA, WSRF and Web Services**

1.3.2.1 Open Grid Service Oriented Architecture (OGSA)

The OGSA broadly defines the capabilities required by services to support grid system. These capabilities are broadly classified under following categories:

a) Execution Management: The Job Execution Management under OGSA addresses issues related to locating executables, selection of execution location, creation of execution environment and initiation and management of the execution.

b) Data Management: Data Management Services are used to move the data, manage replicated copies, run query on data and convert data from one format to another.

c) Resource Management: Managing resource on grid involves resource reservation, monitoring and control.

d) Security Management: These services help in enforcing the security related policy within a Virtual Organization (VO). As Grid specific applications may span multiple administrative domains and each domain may have their own policies and hence grid
applications should adhere to both local-domain policies and the VO policies. The services being developed should be easily able to integrate with the security architecture and should have capability for extension so that the services can easily be integrated with new security services as and when they become available.

e) Self-Management: These capabilities of the services include, self-configuring, self-healing and self-optimizing features to reduce the cost and complexity of owning and operating the resources on grid.

f) Information Services: These services help to manipulate information about applications, resources and other services in a grid environment.

1.3.2.2 Grid Building Blocks

Globus Toolkit is the software developed by Globus Alliance, and can be used for programming grid-based applications. Globus Toolkit consists of a resource monitoring and discovery services, a job submission infrastructure, security infrastructure and data management services. Globus is built as a layered architecture, where lower level services are used to construct higher level of services. The core elements of Globus Toolkit include:

- **GRAM**: Globus Toolkit Resource Allocation Manager
- **GridFTP**: Used for Data access
- **GSI** (Grid Security Infrastructure)
- **MDS** (Monitoring and Discovery Services)
- **Reliable File Transfer**
- **Replica Management**
- **GARA** (Globus Advanced Reservation and Allocation)
GEM (Globus Executable Management)

1.3.2.3 Workflow Management for Grid Computing

Different workflow management software exists to handle complex execution and composition of tasks on grid. The workflow management system helps to build dynamic application; use resources based on user preferences and executes processes across multiple domains by using the available grid Infrastructure.

a) Condor DAGMan: The Directed Acyclic Graph Manager (DAGMan) is a Meta scheduler for Condor jobs. DAGMan uses directed acyclic graphs as a data structures to represent job dependencies, where each job is represented as a node in the graph and edges represents their dependencies. In case of failure of jobs during execution, DAGMan prepares a rescue DAG as an alternate mode of execution.

b) Triana: This is a visual workflow-oriented data analysis environment developed at Cardiff University. Triana (Taylor, et al. 2005) is a java based application, which provides visual programming interface using drag and drop concept. Triana allows users to customize their execution by allowing them to program their own units. These Units can be connected to define the workflow.

c) GridAnt: The GridAnt Workflow management system (Amin, et al. 2004) developed by Argonne National Laboratory provides facility to express and control the execution sequence in a Grid environment. GridAnt consist of four major components, namely workflow engine, run-time environment, workflow vocabulary and workflow monitoring.

d) GrADS: The Grid Application Development Software (GrADS) aims to provide an environment, where ordinary scientific users can program and execute their application on grid.
The above discussed workflow management systems are generally specific to the underlying grid execution environment and are very general in nature. These workflow management systems concentrate on how the applications can be composed, so as to optimally use the resources, but does not cater for specific requirements of earth observation data processing.

Table 1-4 lists out strengths and weaknesses of grid computing architecture.

<table>
<thead>
<tr>
<th>Strengths</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows users to create virtual organization, where resources can be optimally used as shared network enabled infrastructure. The virtual organization can be created from resource anywhere on the globe.</td>
<td>Interoperability between different grid toolkits and standards</td>
</tr>
</tbody>
</table>

1.4 MOTIVATION

In order to cater to user specific processing requirements of earth observation data and to overcome the limitations of transferring high volume of data there is a requirement of a software framework, which allows user to compose an application by defining a workflow using specialized earth observation data processing services and execute them on grid resources co-located with the data. Earth observation data processing software requires capability of plug-and-play and should be modular/scalable/re-usable. These characteristics help in reusing software components in different computing architectures.

Grid based Service Oriented Architecture makes resources available to users on grid as services. Grid services allow software to be integrated into the end-user grid applications in a loosely coupled environment, without knowing module details required for integration. Grid services provides a platform, where the distributed processing
features of grid and capability to expose software as service can be used together, these features are relevant to the current requirements of earth observation data processing as discussed in previous sections. Hence we propose to study and adopt Grid Service Oriented Architecture for earth observation data processing.

Essential components of data processing software includes the core processes, available as executables, which performs the satellite specific corrections and a scheduler, which schedules these processes based on the user specific processing requirements. Earth observation data processing is both computational and I/O intensive, which needs modeling of CPU and I/O loads (Ferrari and Zhou 1988) for scheduling. In case of conventional data processing performance is dependent on the processing power of the system. Computational resources required for performance improvement can be added dynamically using a grid based environment. Grid provides a single interface for requesting and using remote resources for execution of jobs and uses the underlying schedulers namely Condor, OpenPBS and Torque for scheduling of processes. Satellite image processing jobs being CPU and I/O intensive (Li 2005) does not provide the required performance, when integrated under the above-mentioned schedulers. Earth observation data processing also requires the process scheduling service to address functional requirements (Hernández-Torres 2011) such as scheduling in high-availability environment, handling of priority users and event based scheduling. There is a growing need to conduct research on developing scheduling capabilities, which are input data location aware and uses both CPU and I/O loads for scheduling of processes.
As we know data centric processing does not scale well using distributed computing techniques such as Message Passing Interface (MPI) and Parallel Virtual Machine (PVM), as this involves moving a huge volume of data on network.

Earth observation data allows processes to be executed in parallel. Performance can be improved by scheduling jobs on distributed grid nodes, however if a single job has many processes than grid cannot dynamically balance loads, as it cannot pre-empt and migrate the process to other nodes. Distributed Operating System (Sinha 1996) provides a unified view of system under a cluster based processing environment. Distributed operating system allows the processing loads to be balanced by allowing pre-emptive migration of processes to other nodes in a cluster, without any implication of extra programming efforts.

Satellite data selection and its timely dissemination have gained importance in recent years. As discussed earlier earth observation data volumes are growing exponentially. This requires a large network bandwidth for quick dissemination of data to end users. GridFTP (Allcock, et al. 2001) is a standard protocol for accessing data over grid. GridFTP provides capability to transfer data between two remote machines on a grid network (Kourtellis, et al. 2008). Data transfer can be initiated explicitly by the end user or can be delegated to a third machine. The new version of protocol, GridFTP v2 supports several advanced features such as data streaming, dynamic resource allocation, and intermediate transfer, by defining a transfer mode called X-block mode.

Satellite data has different characteristics in terms of data structure/compressibility. So far no research has been done to use these data characteristics in conjunction with the architectural feature including network capability for performance
enhancement; hence it is felt that there is a scope to exploit this unique feature in addition to the existing techniques, which may likely to enhance the performance of GridFTP.

A data transfer model, which freezes parameters before initiation of data transfer and then uses the same during the transfer, is called as static model. If a model dynamically adjusts these parameters during the process of data transfer, without requirement of restarting the session then such a model is called as an adaptive model. The goal of adaptive data transfer is to improve the data transfer performance.

The above issues and requirements has motivated us to carry out extensive research in the area of grid and service oriented architecture including issues of process scheduling for CPU and I/O centric jobs, performance improvement of grid services using distributed operating system and adaptive data transfer using GridFTP protocol.

1.5 REVIEW OF EARTH OBSERVATION DATA PROCESSING ARCHITECTURES

Interoperability is the key issue, when earth observation data is acquired and processed from satellites being operated by different agencies. Efforts are initiated by different space agencies and international organizations such as Committee on Earth Observation Satellites (CEOS) and Group on Earth Observation (GEO) to address these issues to ensure timely processing and dissemination of earth observation data. Grid and Service Oriented Architecture (SOA) appears in the technology road-map of most of the space agencies. NEX (NASA Earth Exchange) allows users to use virtual environment to execute and visualize results. European Space Agency (ESA) G-POD (Grid processing on Demand) (Fusco et al. 2007) is a generic grid infrastructure, which provides capability to the end users to plug in their applications for quick access to data, resources and
results. GEO-GRID (Yamamoto et al 2006) developed by National Institute of Advanced Industrial Science and Technology (AIST), Japan, aims to provide e-Science infrastructure for earth science community. Space Research Institute of NASU-NSAU, Ukraine has developed a grid based system for satellite data processing (Shelestov 2006).

As part of literature survey following research initiatives in the field of earth observation image processing using grid and SOA are reviewed.

a) Information Management for Grid-Based Remote Sensing Problem Solving Environment (Aloisio, et al. 2004): This work aims at design and implementation of a configuration repository for a grid-based problem-solving environment, specialized to describe applications and data belonging to remote sensing. This is a complete integrated computing environment for composing, compiling and execution of applications. This research identifies three kinds of resources: hardware, software and data. Each component has an information model specified, which describes the component itself as part of resource description for discovery. Data distribution is format aware i.e. understands the CEOS format. Applications required for carrying out image processing are classified into three categories i.e. Pre-Processing, Post-Processing and Utility Software. Configuration repository consists of Meta-software schema, which describes about the software and helps the scheduler to submit jobs for execution. Metadata schema, describes input and output data formats. Computational Schema describes about the information required for execution of the application. Finally GUI allows composition of complex applications, using software, metadata and maps them to computational resource in a GRID Enabled environment for execution.
b) *Architecture design of grid GIS and its applications on Image Processing based on LAN* (Shen, et al. 2004): This research work is Geographic Information System (GIS) centric. This work analyzes weakness and problems of traditional GIS, and proposes a method to solve these problems using grid computing and web services. This research work describes differences between Grid Geographic Information System (GIS) and Web GIS and proposes architecture for Grid GIS.

c) IPGE: *Image Processing Grid Environment Using Components and Workflow Techniques*. Image Processing Grid Environment (IPGE) is a project that aims at providing high performance image-processing platform in a grid-computing environment. This is a combination of components and workflow techniques on which complex applications can be modeled as grid workflows.

d) Image Processing for the Grid: A toolkit for Building Grid-enabled Image Processing Applications (Hastings, et al. 2003). This paper presents the design and implementation of a toolkit that allows rapid and efficient development of biomedical image analysis applications in a distributed environment. This toolkit employs the Insight Segmentation and Registration Toolkit (IITK) and Visualization Toolkit (VTK) layered on a component-based framework.

e) Development of Geospatially-enabled Grid technology for Earth Science Applications (Chen, et al. 2009): Open GIS Consortium (OGC) has developed a set of technologies, standards and interface protocols for interoperability of geospatial data, and information systems over web. This paper has integrated the storage and computational power of grid using OGC geospatial services such as resampling, re-projection, geo-rectification and visualization.
Our efforts as part of this research work in on similar lines of adopting grid and service oriented architecture but focuses on customizing the grid and service components and provide them as specialized software services in the form of a software layer build on grid layer (layered architecture) to cater to application specific requirements of earth observation data processing rather than using the existing grid components.

1.6 RESEARCH OUTLINE

This research work proposes customized data processing framework for a class of problem in the area of satellite based optical data processing, which addresses the user specific requirement of processing and dissemination of earth observation data.

1.6.1 Research Objectives

Research objectives are classified under following categories

1. Software architecture and framework for satellite data processing
   a. Study available software architectures, current and future requirements of satellite image processing and propose and validate data processing framework, which fills in the current gaps and can take care of future requirements.

2. Customizations for performance improvement
   a. Improve product throughput and turn-around time by customizing process scheduling services for satellite data processing.
   b. Performance improvement of applications by executing services on a cluster based configuration using Distributed Operating System

3. Data dissemination
a. Capability to define a user area and disseminate data in a format independent mode

b. Modeling based on system, network and image specific parameters to disseminate data with improved performance

1.6.2 Thesis Organization

The thesis is organized into six chapters, major research contributions and achievements in these chapters are discussed as follows

Chapter-1: Provides an introduction to a new paradigm called data intensive science and reviews the concepts of remote sensing and earth observations systems. This chapter further provides overview of activities carried out as part of satellite image processing and discusses the requirements of future earth observation systems. Technology trends being adopted by major space agencies to cater to future earth observation requirements are listed out. Motivations for carrying out this research are discussed in detail. This chapter concludes by providing the outline of research contents. Study results are published in International Conference [3]

Chapter-2: This chapter discusses the architecture for new software framework called GEO-ID (Grid Services for Earth Observation Image Data Processing) developed as part of this research. GEO-ID provides capability for end user to process data of his/her area of interest using software available as services and grid based infrastructure. Suitability of this architecture and research requirements are brought out by carrying out extensive simulations using grid simulation tool called GridSim. This chapter concludes with
discussions on simulation results and proposes the configuration of grid test bed. This work is published in International Journal of Digital Earth [1]

**Chapter-3:** Process Scheduling Service (PSS) is one of the important components of GEO-ID. PSS provides capability to schedule processes associated with the user generated application for processing earth observation data. This chapter discusses the process scheduling algorithms developed as part of this research and further proposes customizations, namely adaptive queue based scheduling, scheduling with remote queues and modeling processing execution sequence as directed acyclic graph. This chapter also addresses the capability developed to cater to time and mission critical needs such as scheduling in high-availability environment, priority and event based scheduling, which is not available as part of standard grid environment. This work related the proposed architecture is published in Journal Indian Cartographer [2]

**Chapter-4:** This chapter discusses the proposed algorithm for load balancing and different process migration techniques. An experiment is conducted where GEO-ID based earth observation applications are executed on distributed operating system. The results are discussed and compared with conventional mode of execution. This chapter further discusses the proposed data handling strategies for performance enhancements of accessing satellite data on a distributed operating system, this include file caching, replication and different access modes available as part of distributed operating system. Important inferences are drawn related to the merit of proposed approaches. This work is published in International Conference [4].
Chapter-5: The performance evaluation of transfer of satellite images using different data transfer mode available in GridFTP is discussed in detail in this chapter. Based on the results of performance evaluation an adaptive data dissemination model is proposed. In addition this chapter also discusses results of the adaptive model for different scenarios of data transfer using the proposed grid test bed.

Chapter-6: This chapter discusses in detail results of execution of three different use case scenarios related to user centric earth observation data processing to validate capability of the proposed software framework. The three application scenarios discussed includes time series data processing, which involves handling of large volume of data (data centric application), real time monitoring of events, where timely dissemination of information is essential (event centric application) and Multi sensor data fusion (compute centric application). Results of execution of these three applications on GEO-ID are discussed. These results are published in International conference on Grid and Cloud Computing [6]

Chapter-7: Thesis ends with a conclusion, which highlights the results of the proposed software framework and the customizations adopted as part of this research work. In addition the merits of the proposed approaches/schemes are brought out with simulations and analysis.

1.7 SUMMARY

Earth observation applications have continuous growth. Increasing data volumes have made data un-movable. End user applications require large computational resources for processing and network bandwidth for quick transfer of data. Current architecture has
limitations such as access to data and software, processing data with user specific requirements, performance issues for time-critical applications and capability to quickly disseminate data. These limitations are bottleneck in efficient utilization of earth observation data by end users. Technology roadmap and research in the field of software and data processing shows that Grid and Service Oriented Architecture have potential to meet the current challenges of user centric data processing and dissemination requirements of earth observation data.

In order to efficiently handle earth observation applications there is a requirement of a software framework and a need to carry out research on the same so as to customize the grid and service components. Literature survey suggest that research needs to be carried on scheduling capabilities of grid, techniques for performance improvement of services using load balancing capabilities in a cluster based configuration and adaptive modelling using satellite image specific parameters in conjunction with the network parameters to improve earth observation data dissemination performance.