On using Grid Computing Technology in the Processing and Archiving Facilities (PAF) of EO Data Providers

Grid Technology for EO applications

The Future of Remote Sensing - October 2006, Antwerp

Bart Beusen[(1)], Geert Borstlap[(1)], Yves Coene [(2)]

[(1)] VITO - Vlaams Instituut voor Technologisch Onderzoek - Flemish Institute for Technological Research, Department of Remote Sensing and Earth Observation Processes, Boeretang 200, B-2400 Mol, Belgium.
Email: bart.beusen@vito.be, geert.borstlap@vito.be
[(2)] Spacebel, I. Vandammestraat 5-7, 1560 Hoeilaart, Belgium.
Email: yves.coene@spacebel.be

ABSTRACT

Earth Observation (EO) applications typically deal with the processing of large data quantities. When looking at the future, new high-resolution sensors, either multi-, super- or hyperspectral, will lead to even higher data quantities to be processed. A Grid service is needed to make sure such large amounts of data can be handled efficiently.

Grids use the resources of many separate computers connected by a network (usually the Internet) to solve large-scale computation problems

The idea behind the recent Grid technology for EO applications is to **move the processing algorithms to the data** instead of having to move large amounts of data to the user's processing site (= time consuming). The processing is done on computers in the data center, which have direct (fast) access to the data. Only the final results need to be transported over the Internet to the end-user. The data center of today evolves to being a processing center.

The users will also benefit from the **data storage** offered by the Grid infrastructure.

Keywords : VITO, CTIV, PAF, Earth Observation, Grid, storage capacity, cpu, computing power, distributed, algorithms, data

INTRODUCTION

Grid

Whereas the World Wide Web is a service for sharing information over the Internet, a Grid is a service for **sharing computing power and data storage capacity over the Internet**. Built on pervasive Internet standards, Grid computing enables organizations to share computing and information resources across department and organizational boundaries in a secure, highly efficient manner [1]. Grid technology is also converging rapidly with widely spread Web service standards such as SOAP, WSDL which will accelerate and facilitate its adoption.

Grid computing takes advantage of many networked computers to model a virtual computer architecture that is able to distribute process execution across a parallel infrastructure. Grids use the resources of many separate computers connected by a network (usually the Internet) to solve large-scale computation

problems. Grids provide the ability to perform computations on large data sets, by breaking them down into many smaller ones, or provide the ability to perform many more computations at once than would be possible on a single computer, by modeling a parallel division of labour between processes [2].

Grid technology today still is a "work in progress", with the underlying technology in a development phase. However, increased deployment of Grid technologies in the commercial sector over the last years seems to suggest that Grids are **on the verge of a breakthrough**.

Organizations around the world are utilizing Grid computing today in such diverse areas as high-energy physics, human genome research, drug discovery, financial risk analysis, product design, and … Earth Observation. Indeed, the European Space Agency (ESA) has also initiated Earth Observation – related Grid initiatives such as Grid On Demand (GODIS) [7] and SpaceGrid.

EO applications - PAF

EO data providers make use of a Processing and Archiving Facility (PAF) to gather raw sensor image data, process it up to a certain level (applying geometric correction, atmospheric correction, cloud detection, 10-day syntheses, ...) and archive the results. Users can order existing products through the catalog service of the EO provider. Using a system of subscriptions, users may automatically download new products as soon as they are available.

In practice, the architecture and functionality of a traditional PAF imply that archives are continuously being duplicated and large amounts of data are being moved around over limited Internet lines.

DRAWBACKS FOR EO PROCESSING

The current situation for the users of basic Vegetation (VGT) products, as they are offered at this moment by the CTIV (*Centre de Traitement des Images VEGETATION*) at VITO, is shown in Figure 1: the user first places an order for certain basic VGT product, he then receives the information necessary to download these products. He can use the downloaded data to run his algorithms and produce his results locally on his own computer infrastructure.

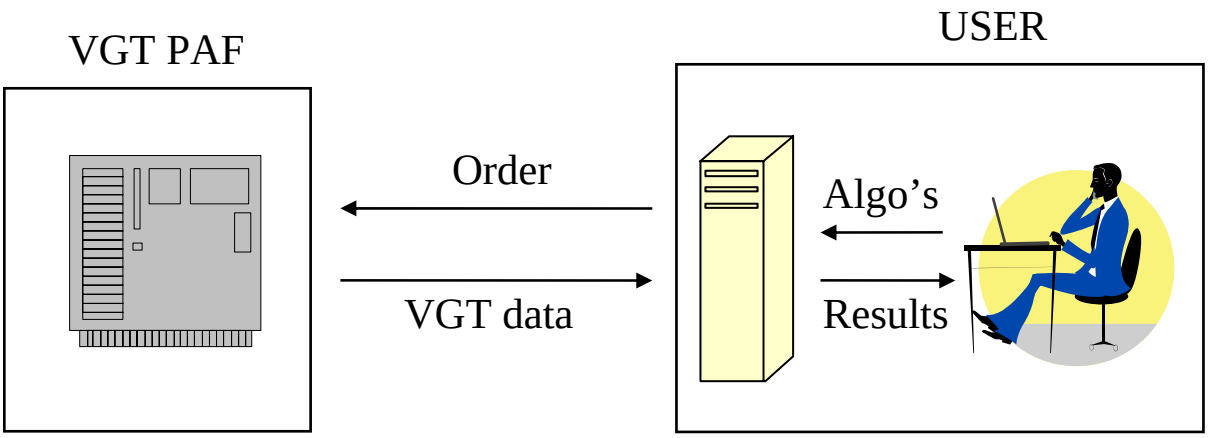

Figure 1 : Current situation for the VGT-data user

Typically for EO applications, the size of the products can be very large, and the data (possibly from different sensors) might be stored in different institutions. This can result in **very long wait times** to download the data in order to further process it. This situation also puts high demands (thus high costs) on the user's **storage capacity**.

When looking at the future, new high-resolution sensors, either multi-, super- or hyperspectral, will lead to even higher data quantities to be processed. A Grid service is needed to process such large amounts of data in an efficient way.

As an example, data specifications for a basic "S10" VGT product are presented below. The S10 product is a synthesis product over a ten day period, delivered as a zip-file, containing 11 bands (B0, B2, B3, MIR, NDV, SAA, SM, SZA, TG, VAA and VZA) in separate images [3]. Typical file sizes for the S10 product are listed in Table 1 below.

Table 1 : data quantities for a CTIV S10 product

Product	Size
S10 zipped	1,29 Gb
S10 unzipped	8,9 Gb
B0, B2, B3, MIR, TG	1,1 Gb
NDV, SAA, SM, SZA, VAA, VZA	564 Mb

In the case of time series analysis, multiple S10's need to be downloaded and unpacked. In the case an analysis is required on 1 year of ten-daily VGT composits, the user will need to provide 320 Gb data storage. Apart from the duplication of large amounts of data, and bandwidth requirements for download, this situation also imposes a large cost on data storage.

GRID SOLUTION FOR EO APPLICATIONS

The idea behind the recent Grid technology is to **bring the processing algorithms to the data** instead of having to move large amounts of data to the user's processing site. The processing is done on computers in the data center, which have direct (fast) access to the data. Only the final results need to be transported over the internet to the end-user. In most cases, the size of the result is smaller the size of the input.

The users will also benefit from the **data storage** offered by the Grid infrastructure.

Users will have **direct access to EO input images,** since the processors to run the algorithms have direct access to the fileservers containing the EO data.

Furthermore, the users have the possibility to **use existing software modules** made available to them by the EO provider, or they can run their own algorithms, or they can even use a combination of both. Grid enables end-users to run and control their processing chains locally, while using a centralized processing infrastructure which is directly connected to the data archives.

VITO and Spacebel plan to provide access to Grid services (on an ESA and VITO Grid) via the ESA Service Support Environment (SSE) [6]. This future SSE-Grid integration is shown in Figure 2: the user can upload his algorithms to the Grid-infrastructure of the EO provider, process the sensor data using his custom defined workflow, and then download the results. An initial prototype making available an IMERIS GODIS Grid processor on the ESA Grid and allowing invoking it via the SSE was demonstrated earlier this year as part of the ESIT project. The proposed provision of processing interfaces on the PAF is also fully in-line with the processing use cases and scenarios defined by the ESA Heterogeneous Missions Accessibility (HMA) GMES preparatory study which is currently managed by ESA/ESRIN in cooperation with the major European Space Agencies.

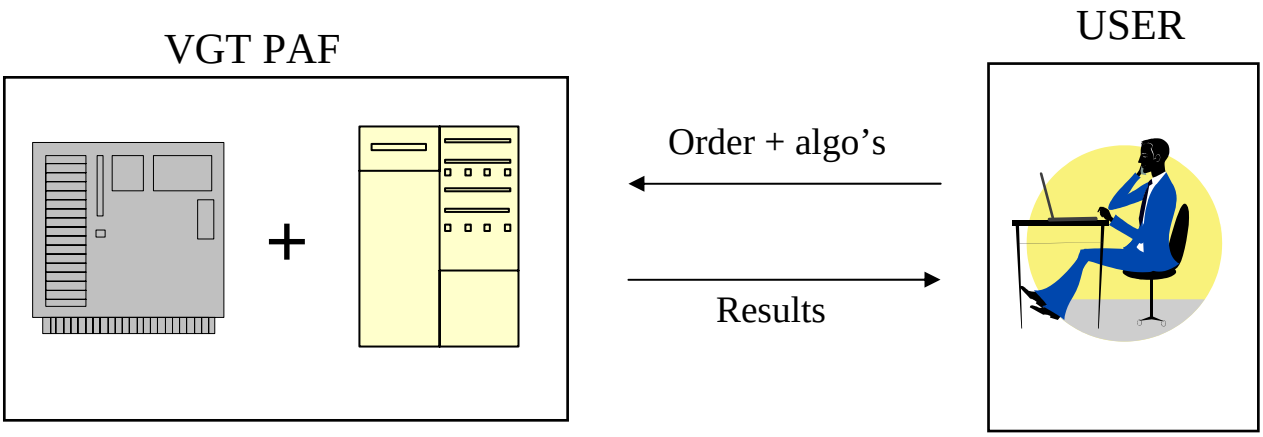

Figure 2 : Future situation for the SSE-Grid user

Data from different sensors is mostly stored at different geographic locations. For example the VGT images are processed at the CTIV PAF at VITO in Mol, Belgium, whereas ATSR2/AATSR images are processed at the ESA/ESRIN PAF in Frascati, Italy. If both PAF's are connected in one Grid, the user can run his processing chain on either dataset, without having to worry about the physical location of the data. The Grid will decide where the processing is performed, typically as close to the data as possible. The user will even have the possibility to combine both VGT and AATSR data in one chain, while the Grid middleware takes care of the optimal processing configuration.

The advantages of a Grid for EO processing are clear :

- no more need to copy large amounts of data (time consuming) from provider to the user locally
- powerful processing units , leading to shorter processing times,
  Benefit : The processing capacity can be shared amongst users.
- a large data storage capacity with direct access to the sensor data,
  Benefit : This avoids a heavy consumption of bandwidth by moving data around.
- Giving scientists the possibility to run algorithms on the data by submitting the algorithm on a Grid without them having to obtain access to the actual data or download the data, e.g. via an ESA CAT-1 request decreases significantly the data policy and license (digital rights management) issues.
- Offer the use of existing software components (e.g. atmospheric correction, cloud detection, etc.) by making them available in the Grid environment. These pre-built components can be directly used in a users own custom processing chain.
- Possibility for a "tape service" (implemented as a Web service), so that processed results can be written on tape or on DVD and ordered at nominal costs.

DIFFERENCE WITH OTHER GRID ENVIRONMENTS

The PAF of an EO provider gathers raw sensor image data and processes it up to a certain level. The processing of one particular segment is done in several consecutive steps which cannot be run in parallel. Each step generates intermediate results, which are needed in the subsequent steps. The processing of several segments at the same time can however be run in parallel.

As a consequence, data access (disk IO) is often the limiting factor for the processing of EO images, rather than CPU power. Disk access will need to be as fast as possible for the processing to be efficient.

Therefore, the processing of EO images is best done on worker nodes which have direct access to the data storage, being the worker nodes which reside in the same Grid-cluster as the file servers. Disks can be mounted locally on the worker nodes using e.g. the NFS protocol.

This results in a PAF which has following characteristics (example given for the CTIV PAF) [4]

- storage capacity : 120 terabytes
- computing elements : 45

Whereas a traditional Grid infrastructure is focused primarily on computing power, and less on terabytes of storage capacity. The infrastructure of the BeGrid [5] (Belgian Grid infrastructure that results from the Belnet Grid Initiative) today consists of:

- storage capacity : 3 terabytes
- computing elements : > 400

It is clear from the example given above, that an EO Grid infrastructure requires a huge amount of storage capacity when compared to a traditional Grid infrastructure, and puts relatively less demands on computing power capacity.

INNOVATIVE SOLUTIONS AT VITO

Near-real time processing

Typically, Grid technology in EO applications is used for the batch processing of historic data. The Grid project which VITO is working on, will offer **processing in near-real-time**. Users will have the possibility of creating their own processing and archive facility, **running their own algorithms and workflows**. As soon as new satellite input data arrives, their algorithms could be triggered into action.

Similar to a subscription in a traditional PAF, where generated products can be downloaded as soon as they become available, the near-real time processing allows users to apply their own processing chain on the raw incoming sensor images as soon as they are available. When the processing is done, the user is notified so he/she can download the fresh results.

Data delivery service for the Grid

In most cases, the analysis results in EO applications is significantly smaller than the input data. However, for some applications the output can still be very large, sometimes even comparable to the amount of input data. Getting the results to the user in a fast and secure way must be guaranteed. A **tape-write service**, or **DVD-burn service**, offers the possibility of delivering a lot of data quickly and efficiently by mail.

CONCLUSIONS

EO data providers make use of a PAF (Processing and Archiving Facility) to gather raw sensor image data, process it up to a certain level (applying geometric correction, atmospheric correction, cloud detection, 10-day syntheses, ...) and archive the results. Users can order processed products through the catalog service of the EO provider. Data quantities are very large and fast and direct access to the input data is needed for efficient processing.

The idea behind the recent Grid technology is to **bring the processing algorithms to the data** instead of having to move large amounts of data to the user's processing site (= time consuming). The processing is done on computers in the data center, which have direct (fast) access to the data. Only the final results,

which are mostly much smaller in size than the input data, needs to be transported over the Internet to the end-user. Furthermore, the user benefits from the computational power and from the data storage capacity of the Grid-infrastructure.

REFERENCES

[1] "Open Grid Forum" - http://www.ogf.org/
[2] http://en.wikipedia.org/wiki/Grid_computing
[3] http://www.spot-vegetation.com/vegetationprogramme/index.htm
[4] VITO (CTIV) - CVB Informatica Infrastructure
[5] http://www.begrid.be/
[6] http://services.eoportal.org
[7] GODIS, http://giserver.esrin.esa.int/cgi-dev/ws/eogrid.cgi