Image processing methods and architectures in diagnostic pathology

Gloria Bueno[1], Oscar Déniz[1], Jesús Salido[1], Marcial García Rojo[2]

[1]Engineering School, Universidad de Castilla-La Mancha, Ciudad Real, Spain
[2]Pathology Department, Hospital General de Ciudad Real, Ciudad Real, Spain

Abstract: Grid technology has enabled the clustering and the efficient and secure access to and interaction among a wide variety of geographically distributed resources such as: supercomputers, storage systems, data sources, instruments and special devices and services. Their main applications include large-scale computational and data intensive problems in science and engineering. General grid structures and methodologies for both software and hardware in image analysis for virtual tissue-based diagnosis has been considered in this paper. This methods are focus on the user level middleware. The article describes the distributed programming system developed by the authors for virtual slide analysis in diagnostic pathology. The system supports different image analysis operations commonly done in anatomical pathology and it takes into account secured aspects and specialized infrastructures with high level services designed to meet application requirements. Grids are likely to have a deep impact on health related applications, and therefore they seem to be suitable for tissue-based diagnosis too. The implemented system is a joint application that mixes both Web and Grid Service Architecture around a distributed architecture for image processing. It has shown to be a successful solution to analyze a big and heterogeneous group of histological images under architecture of massively parallel processors using message passing and non-shared memory.

Key words: Image processing, grid architecture, digital slide image analysis, diagnostic pathology.

Introduction

The Grid vision has been described as a world in which computational power is as readily available as different utilities [1]. They are available to users by means of computational, data, application, information and knowledge services at different levels and areas [2]. Grid technologies have been developed for applications with large storage and computation requirements [3]. They offer a powerful tool to deal with current challenges in many biomedical domains involving complex anatomical and physiological modelling of structures from images or large image databases assembling and analysis. Indeed, grids are suitable for: a) allowing distribution of large datasets over different sites, b) enforcing the use of common standards for data exchanges, c) enlarging the datasets available for large scale studies in many sites, d) opening new studies, perspectives and application fields.

On the other hand, institutions involved in tissue-based diagnosis, including conventional, prospective, indicative and risk-assigned diagnosis, should have access to a variety of sources for data, information, and knowledge, to enable working in an efficient manner [4]. That is why Grids are a promising technology for Diagnostic Pathology. Particularly, medical image processing and the introduction of virtual slides will promote Grid applications in tissue-based diagnosis to fulfil the requirements of Anatomical Pathology applications.

Grid computing has been applied in several large-scale projects in biomedicine, such as: BIRN, MammoGrid, eDiamond, MEDIGRID, caBIG™ and EAMUS™. The Biomedical Informatics Research Network (BIRN) [5] initiative focuses on support for collaborative access to and analysis of datasets generated by neuroimaging studies. Both MammoGrid [6] and eDiamond [7] projects are funded by the European Union (EU) to apply Grid user level middleware and network resources to build and research on a distributed database of mammograms. These projects develop and promote standardization in medical image databases for mammography and other cancer diseases.

Correspondence: G. Bueno, Universidad de Castilla-La Mancha, E.T.S. Ingenieros Industriales, Avda. Camilo José Cela, s/n 13071 Ciudad Real, Spain; e-mail: gloria.bueno@uclm.es

MEDIGRID [8] is another multi-institutional project investigating the application of Grid technologies for manipulating large medical image databases.

In Anatomical Pathology, the caBIG™ (cancer Biomedical Informatics Grid) project [9] released in 2006 tries to develop applications and the underlying systems architecture that connects together data, tools, scientists and organizations in an open federated environment. caBIG™ used caGrid as the underlying SOA (service oriented network architecture) that provides the basis for connectivity between all of the cancer community institutions. caTissue is a project developed in 2007 and integrated within caBIG™ to develop software repository tools for biospecimen inventory, tracking and basic Pathology Annotation (CAP protocol pathology annotation for major organ systems), caTIES-like Pathology Report Annotation, Custom Annotation (Dynamic Extensions) and Advanced Query "Wizard". caTissue allows users to track the collection, storage, quality assurance, and distribution of specimens as well as the derivation of new specimens from existing ones (*e.g.* for DNA analysis). It also allows users to find and request specimens that may then be used in molecular, correlative studies. The objectives of caTissue are a) establishing Common Data Elements b) establishing protocols for the new tracking system and data fields c) establishing parity for conversion from legacy to the new tracking system and d) adoption of caBIG™ compatible caTissue systems. The tool, when refined, will enhance the ability of pathologists, biospecimen bankers and other researchers to share databases and locate and acquire specimens with the long term goal of catalyzing translational research and optimizing biospecimen banking efforts [9].

In tissue-based diagnosis, the EAMUS™ project [10] can be considered as a simple, one-node implementation of the application service. The application services manage Grid application and give access to remote software, libraries and Web services [4]. Moreover, the EAMUS™ project may be used for a grid information service by combining it with an existing telepathology information system such as UICCTPCC, or iPATH [11]. Information services try to extract and present information provided by data and/or other grid services, and to put these into relationship. Other example of a grid information service is the EU project Virolab Grid. This project deals with viral infection (HIV/AIDS) analysis. Its service offers the integration of biomedical information, advanced applications and data services under secure distributed resources [4,12].

Other grid solutions in diagnostic pathology for data and computational services have been developed. That is: a) to offer secure access to distributed datasets, managing access, retrieval, storage, replication, or catalogues of individual or distributed libraries, in the case of data services, and b) to deal with secure distributed computational resources for executing application jobs in the case of computational services. These solutions are based on Picture Archive and Communication System applications developed for live imaging, neurosurgery or dermatology [13,15], evaluation of cytology smears [16], or DNA measurement or expression of antigens [17,18]. Also the application of the workflow system DataCutter [19] for efficient image processing workflows on very large microscopy image datasets across Grid-nodes have been used for mouse brain analysis [20] and prognosis of neuroblastoma [21].

These examples show how Grids have emerged as a promising technology to handle large amounts of data and compute the specific biomedical and bioinformatics requirements for several applications in radiology, neurosurgery, etc. However, grid technologies are still in their youth and they often propose only very generic services for secure access. Higher level services should be developed to take into account the specific requirements in biomedical applications and particularly in diagnostic pathology. The above mentioned tools can only be considered as precursors and do not meet the performance of a Grid in general, as they are designed for one analysis system with open access. Implementations of Grids to be applied in tissue-based diagnosis have not been published yet [4].

The following sections contain a brief introduction to a general workflow scheme in Diagnostic Pathology connecting this to grid architecture and individual components in distributed tissue-based diagnosis. Then the characteristic for a Grid Distributed Architecture will be described. The paper ends with the explanation of the grid middleware application implemented by the authors for Diagnostic Pathology applications.

Materials and methods

Diagnostic pathology as a grid application. The workflow in an institute of pathology usually follows a time sequence order of tissue identification, slide production, processing, patient – oriented diagnostics, secretary work and final diagnosis submission. The described workflow is not a simple straightforward procedure but there might be several feedback mechanisms due for example to slide quality and additional clinical information. Moreover, emerging technologies in anatomical pathology imaging, that is the introduction of virtual microscopy, will cause several changes on the workflow of the pathology institute. Mainly it will offer the chance to interact with all feedback mechanisms directly. For example, additional data, clinical information and second opinions can be requested prior to diagnosis and report. The actual pathology information system will also change to allow storage and retrieval of the digital slides, as well as include diagnosis databases linked to the images [22, 23]. According to the workflow, several grid functions may be distinguished in at least five different grid levels [24]. These levels are:

- Application Level: Laboratory and diagnostic quality assurance, slide screening, measurements, digital diagnosis reports, expert consultation and clinical history archive.

- User Level Middle Ware: Software environment and tools (programming languages, libraries, etc…), diagnosis management and data selection.
- Core Level Middle Ware: Pathology – hospital, institutions coupling services. Private security, images and clinical data and diagnosis status.
- Local resources management: Laboratory, pathology operating and queuing systems, libraries, hospital archive, internet protocol.
- Network resources (hardware): Computers, line connections, storage systems, scanners, monitors, tissue embedding and slide preparation instruments.

Distributed architecture. Distributed programming models and systems are the basis of most existing Grid programming and applications. They allow a very efficient environment to integrated different type of data and algorithms [2]. Their characterization, requirements and classifications are as follows.

Characteristics of Grid execution systems and applications. There are four main characteristics of Grid execution systems and applications:

- Heterogeneity: Grid environments aggregate large number of independent and geographically distributed computational and information resources.
- Dynamism: The grid computation, communications and information environment is continuously changing during the lifetime of an application. Moreover, the organization and interactions of the component/services can also change.
- Uncertainty: caused by multiple factors, including: a) Dynamism, which introduces unpredictable and changing behaviors, b) Failures, which have an increasing probability of occurrence as systems/application scales increase, c) Incomplete knowledge of global system state, which is intrinsic to large distributed environments
- Security: Grid has secure hardware/software resources for authentication, authorization and access control.

Requirements for Grid systems. The above mentioned characteristics impose grid systems to be able to specify applications which can detect and respond to changes in both the execution and application states. That is:

- Grid applications should be composed from discrete, self-managing components which incorporate separate specifications for all states:
  a) functional: computational,
  b) non-functional: performance, fault detection and recovery,
  c) interaction and coordination.
- The specifications of the different states or behaviors should be separated and their combinations should be re-composed.
- The interface definitions of these components should be separated from their implementation to enable heterogeneous components to interact and to enable dynamic selection of components.

Classification for Grid systems. Programming models for grid systems may be classified as:

1) Models based on the addition of communication/interaction models and mechanism to sequential programming models, that is: Message Passing Interface (MPI), Parallel Virtual Machine (PVM), shared-space models and RPC (remote procedure calls).
2) Distributed object models, for example Common Object Request Broker Architecture (CORBA).
3) Component based models such as: JavaBean, CORBA Component Model (CCM), Common Component Architecture (CCA).
4) Services models, for example: Web Service, Open Grid Services Architecture (OGSA) and Web Service Resource Framework (WSRF).

Results

In this paper we deal with an application for the User Level Middleware. The objective of the application is the implementation of a Web based application accessible, efficient and useful in diagnosis pathology to visualize and analyze different type of image formats (jpeg, jpg, tif, gif, ppm, bmp) at different zoom levels. The application, obviously, must also take into account that the size of a pathology image is typically in the order of several gigabytes. At the same time the application should give support to medical requirements, that is integrate collaborative tools, editing and navigation tools as well as image pre-processing and processing methods.

The distributed programming application implemented by the authors is a joint application that mixes both Web and Grid Service Architecture around a distributed architecture for image processing. It makes use of servlets-Java and Java Server Pages, as well as models based on MPI and MPI2. The application gives support to multiple users by means of a server (Apache Tomcat) and it may be integrated within a net-platform. Fig. 1 shows the scheme to integrate the web services within a network platform.

The service based models (Web Service) exist across applications against component based models that has meaning during the life time and in the context of an application. Web services are easy to create, to access and to integrate different data and programming models. Message Passing models /Interface (MPI) provide messaging abstraction that enable entities defined by sequential programming models to communicate; that is message passing operators to sequential languages such as C and Fortran. MPICH-G2 and MPI-2, a grid enabled implementation of MPI, support heterogeneity using services provided by Globus toolkit, dynamism and uncertainty.

For the system design unified process and conceptual design (UML) was used for case diagrams. Also a logical design for services and client objects by means of UML diagrams was done and divided into 3 layers:

- Functional: computational domain.
- Non-functional: performance, output processing, etc.
- Interaction-coordination: presentation for the Graphical User Interface (GUI).

Architecturally, the application has been divided into three main parts:

- Visualization: virtual microscopy viewer, using: a) Applet (Java): Hypertext Transfer Protocol (HTTP) over TCP at port 80, security for external institutions, b) HTTP Tunneling: applet communication with remote servers, c) HTTP Basic Authentication of client (login and password), d) Java library for GUI, data-file interchange and pre-processing.
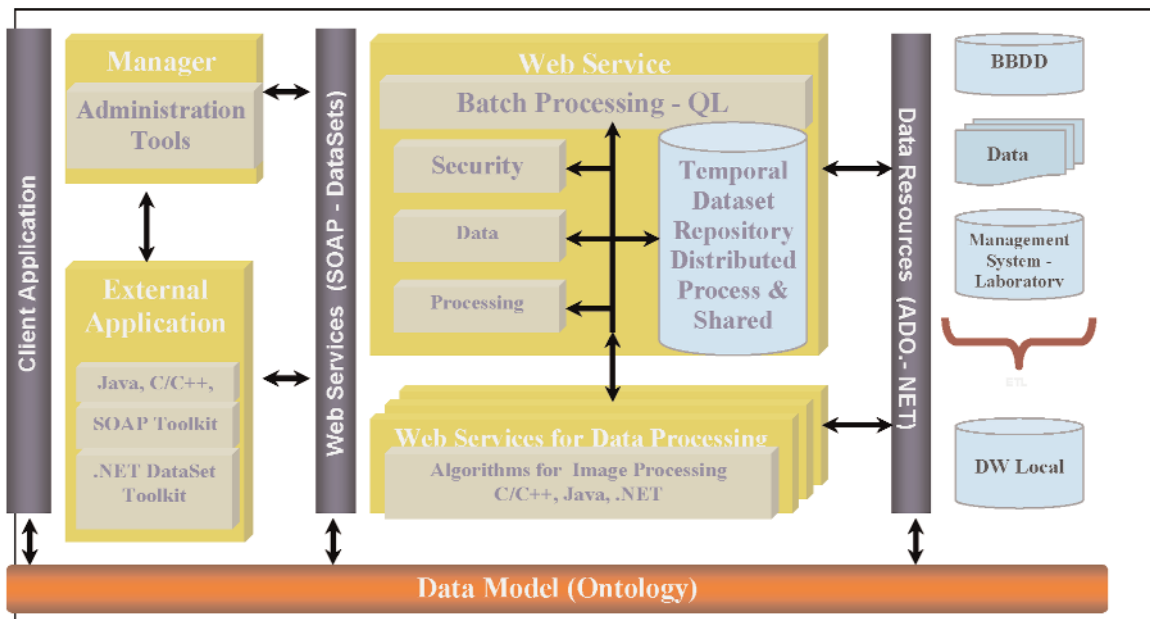
Fig. 1. Web-Grid Service Architecture Integrated within a Network for Platform.

- Processing: web servlets processing, using: a) C, C++ for an efficient parallel image processing, b) NET: Platform integration for the web application, c) General platform IQL, for example INBIOMED (25).
- Storage: data storage, using: Servlets technology, Java 2 Enterprise Edition (J2EE) and Java Server Pages (JSP), and looking for efficiency when manipulating the gigabyte size images.

An MPP architecture with a cluster of 17 nodes interconnected using INFINIBAND, each node having one processor Intel XEON of 3.2GHz and 2GB of RAM memory was the hardware configuration used.

Fig. 2 shows working sessions of the implemented system for diagnostic pathology. The systems support visualization and editing tools for all pathologic images and processing for just biopsy, cytology and autopsy histological images. Fig. 2a shows a cytology image for processing and measurement of regions of interest (ROIs). The system also allows different editing tools for collaborative work to introduce comments, questions and answers about ROIs within the image. Fig. 2b shows the ROIs detection and diagnosis carried out by the system after the application of artificial intelligence techniques. The Fig. shows the processing of a whole cytology virtual slide is shown together with the results of the ROIs diagnosis and the localization of those ROIs.

A parallel model for the processing of the virtual slides based on MPI-2 under suitable colour spaces and distances has been developed [26]. Additional results of the parallel processing on the tissue samples used in this research are illustrated in Fig. 3 with fragments of 100 Mbytes. Figs. 3 a, b and c shows the parallel segmenta-

tion of a fragment of prostate biopsy, cytology and a brain autopsy at 40x in its mesencephalic nucleus.

In order to see the benefits of developing parallel image processing algorithms from the efficiency viewpoint different tests were carried out for the segmentation of ROIs from lung cytology of 200 Mbytes under several processors. The colour model used was the CIEL*a*b* with the CIEDE2000 colour distance formula. The results for a fixed number of ROIs and increasing the number of processor is shown in Fig. 4. Finally Fig. 5 shows the cost of segmenting with CIEL*a*b* and HIS colour model applied to different image sizes and using 17 nodes. It is possible to see how time decreases exponentially with the number of processors decrease exponentially. Moreover, the time of the parallel processing is quite dependent of the colour segmentation algorithm, as shown in Fig. 5.

The implemented parallel model is a successful solution to parallelize a big and heterogeneous group of image processing algorithms under architecture of massively parallel processors using message passing and non-shared memory.

Discussion

Grids have emerged as a promising technology to handle large amounts of data and compute the specific biomedical requirements in different areas such as radiology, bioinformatics and neurosurgery. However, further efforts should be made to implement Grids for tissue-based diagnosis. These efforts will be fuelled with the use of virtual microscopes.

Only secured and specialized infrastructures with high level services including automated diagnostic
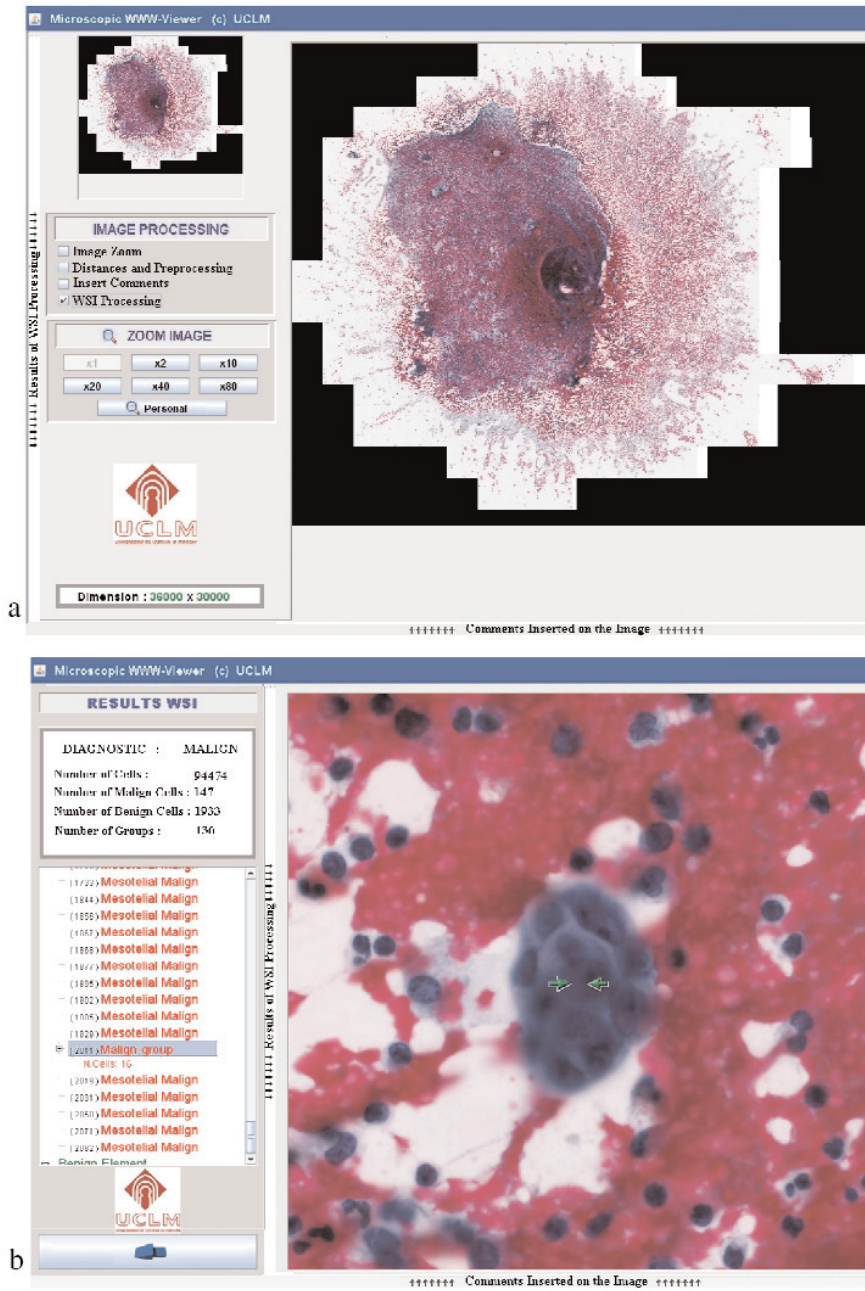
Fig. 2. Working sessions of the implemented Web-Grid system for diagnostic pathology. **a**. Virtual slide visualization and processing, **b**. ROI detection and diagnosis.
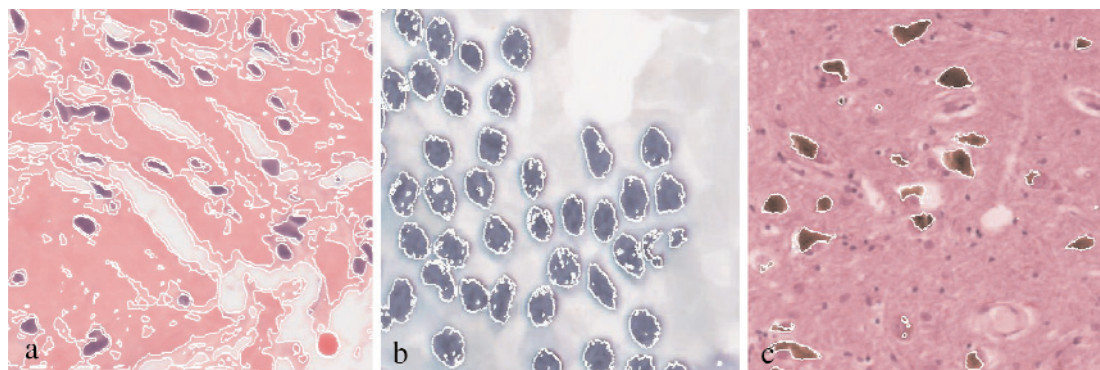

Fig. 3. Whole slide parallel image segmentation of histological images. **a**. Biopsy. **b**. Citology. **c**. Autopsy virtual slides.
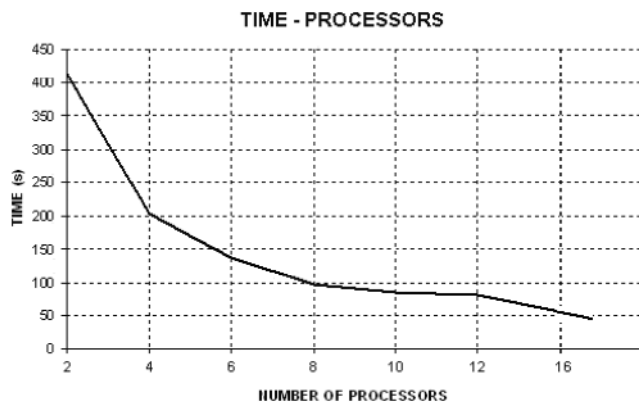
Fig. 4. Computational cost analysis based on the number of processors.

procedures and appropriate standards will really allow a large scale acceptance and deployment of grid environments for diagnostic pathology applications.

This paper has given a general view of Grid distributed programming models, their characteristic and requirements applied to diagnostic pathology. It has been shown how grids have been used for biomedical applications. Moreover, basic middleware is today available and there is a big potential in grid technologies to tailor specific applications.

A Grid middleware application for diagnostic pathology, implemented by the authors, has been presented. The system supports different image analysis operations commonly done in anatomical pathology. The system takes into account secured aspects and specialized infrastructures with high level services designed to meet application requirements.

The complexity, characteristics and data dependencies of histological images are very high and many constraints exist that do not allow creating a completely automatic parallel image processing tool. Each tissue and algorithm must be analyzed to define the targets and detect data dependencies to address their solutions. The implemented MPI image processing software has shown to be efficient and to have a high performance. In order to make an optimal parallel implementation taking into account all data dependencies and the relationship between number of processors, CPU time at each node and network traffic, a previous study of the image processing algorithm is required for each application.
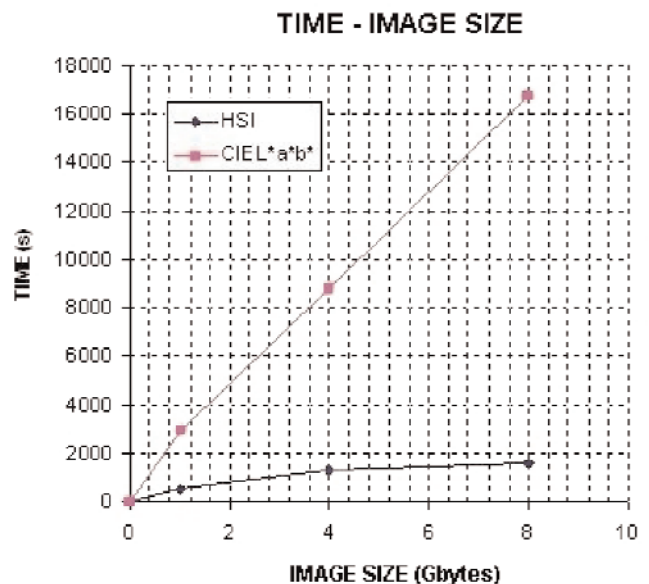
Fig. 5. Computational cost analysis based on the segmentation time for different image sizes.

References

[1] Roure D, Jennings NR, Shadbolt NR. The semantic grid: past, present and future. *Proceedings of the IEEE.* 2005;93(3):669-681.

[2] Parashar M, Browne JC. Conceptual and implementation models for the grid. *Proceedings of the IEEE.* 2005;93(3): 653-668.

[3] Montagnat J, Breton V, Magnin IE. Using grid technologies to face medical image analysis challenges. In: *Proceedings of BioGrid* 2003: 3rd Int. Symp. Cluster Comput. Grid (CCGrid 2003), pp. 588-593.

[4] Görtler J, Berghoff M, Kayser G, Kayser K. Grid technology in tissue-based diagnosis: fundamentals and potential developments. *Diagn Pathol.* 24 August 2006, 1:23, from: http://www.diagnosticpathology.org/content/1/1/23.

[5] Peltier S, Ellisman M. The Grid, blueprint for a new computing infrastructure. In: *The Biomedical Informatics Research Network*, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2003, from: http://www.nbirn.net.

[6] Amendolia R, Estrella F, Hauer T, Manset D, McCabe D, McClatchey R, Odeh M, Reading T, Rogulin D, Schottlander D, Solomonides T. Grid databases for shared image analysis in the mammogrid project. In: *Proc. 8th Int. Database Eng. Appl. Symp.* (Ideas 2004), pp. 302-311.

[7] Solomonides A, McClatchey R, Odeh M, Brady M, Mulet-Parada M, Schottlander D, Amendolia S. Mammogrid and ediamond: Grids applications in mammogram analysis. In: *Proc. International Association for Development of the Information Society International Conference, IADIS Int. Conf.*: e-Soc. 2003, pp. 1032-1033.

[8] Tweed T, Miguet S. (2003). Medical image database on the grid: Strategies for data distribution. In: *Proc. HealthGrid.* 2003 Workshop, pp. 152-162.

[9] caBIG, caTissue. (2006-2007). From: https://cabig.nci.nih.gov/tools/catissuecore, and https://cabig.nci.nih.gov/workspaces/Architecture/caGrid http://gforge.nci.nih.gov/projects/cacorecvs/

[10] Kayser G., Radziszowski D., Bzdyl P., Sommer R., & Kayser K. Theory and implementation of an electronic, automated measurement system for images obtained from immunohisto-

chemically stained slides. *Anal Quant Cytol Histol*. 2006: 28(1);27-38.

[11] Kayser K, Molnar B, Weinstein RS. Virtual Microscopy. Published by VSV Interdisciplinary Medical Publishing, 2006.

[12] Oberholzer M, Christen H, Haroske G, Helfrich M, Oberli H, Jundt G, Stauch G, Mihatsch M, Brauchli K. Modern telepathology: a distributed system with open standards. *Curr Probl Dermatol*. 2003:32;102-14.

[13] Virolab. (2006). From: http://www.virolab.org:8080/virolab/ and http://www.gridwisetech.com/virolab.

[14] Germain C, Breton V, Clarysse P, Gaudeau Y, Glatard T, Jeannot E, Legre Y, Loomis C, Magnin I, Montagnat J, Moureaux JM, Osorio A, Pennec X, Texier R. Grid-enabling medical image analysis. *J Clin Monit Comput*. 2005:19(4-5);339-349.

[15] Liu BJ, Zhou MZ, Documet J. Utilizing data grid architecture for the backup and recovery of clinical image data. *Comput Med Imaging Graph*. 2005:29(2-3);95-102.

[16] Balogh AA, Preul MC, Laszlo K, Schornak M, Hickman M, Deshmukh P, Spetzler RF. Multilayer image grid reconstruction technology: four-dimensional interactive image reconstruction of microsurgical neuroanatomic dissections. *Neurosurgery*. 2006:58(1 Suppl):ONS157-165.

[17] Husain OA, Watts KC. Preparatory methods for DNA hydrolysis, cytochemistry, immunocytochemistry and ploidy analysis. Their application to automated and routine diagnostic cytopathology. *Anal Quant Cytol Histol*. 1987:9(3);218-224.

[18] Haroske G,Meyer W, Kunze D, Boeking A. Quality control measures for dna image cytometry in a telepathology network. *Adv Clin Path*. 1998:2(2);143-145.

[19] Haroske G, Meyer W, Oberholzer M, Bocking A, Kunze KD. Competence on demand in DNA image cytometry. *Pathol Res Pract*. (2000:196(5);285-291.

[20] Beynon MD, Kurc T, Catalyurek U, Chang C, Sussman A, Saltz J. Distributed processing of very large datasets with DataCutter. *Parallel Computing*. 2001:27(11);1457-1478.

[21] Kumar VS, Rutt B, Kurc T, Catalyurek UV, Pan TC, Chow S, Lamont S, Martone M, Saltz JH. Large-Scale Biomedical Image Analysis in Grid Environments. *IEEE Trans. on Information Technology in Biomedicine*. 2008:12(2);154-161.

[22] Cambazoglu B, Sertel O, Kong J. Efficient processing of pathological images using the grid: Computer-aided prognosis of neuroblastoma. In: *Proc. Challenges of Large Applications in Distributed Environments*, CLADE'07, June 25, 2007, pp. 35-41, Monterey, California.

[23] Schrader T, Niepage S, Leuthold T, Saeger K, Schluns K, Hufnagl P, Kayser K, Dietel M. The diagnostic path, a useful visualisation tool in virtual microscopy. Diagnostic Pathology. doi:10.1186/1746-1596-1-40, 08 November 2006, 1:40, from: http://www.diagnosticpathology.org/content/1/1/40.

[24] Kayser K, Görtler J, Goldmann T, Vollmer E, Hufnagl P, Kayser G. Image standards in tissue-based diagnosis (diagnostic surgical pathology). *Diagn Pathol*. 2008;3:17 from: http://www.diagnosticpathology.org/content/3/1/17.

[25] Pérez del Rey D, Crespo J, Pérez LJ, Dorado J, Bueno G, Feliu V, Estruch A, Heredia JA. Biomedical Image Processing Integration through INBIOMED: A Web Services-Based Platform. Lecture Notes in Computer Science, (LNBI). 2005; 3745:34-43.

[26] Bueno G,González R, Déniz O, González J, García-Rojo M. Colour model analysis for microscopic image processing. *Diagn Pathol*. 2008;3(Suppl 1):S18.