Cyber-Infrastructure for Data-Intensive Geospatial Computing

Rajasekar Karthik, Alexandre Sorokine, Dilip R. Patlolla, Cheng Liu, Shweta M. Gupte, and Budhendra L. Bhaduri

With the recent advent of heterogeneous High-performance Computing (HPC) to handle EO “Big Data” workloads, there is a need for a unified Cyber-infrastructure (CI) platform that can bridge the best of many HPC worlds. In this chapter, we discuss such a CI platform being developed at Geographic Information Science and Technology (GIST) group using novel and innovative techniques, and emerging technologies that are scalable to large-scale supercomputers. The CI platform utilizes a wide variety of computing such as GPGPU, distributed, real-time and cluster computing, which are being brought together architecturally to enable data-driven analysis, scientific understanding of earth system models, and research collaboration. This development addresses the need for close integration of EO and other geospatial information in the face of growing volumes of the data, and facilitates spatio-temporal analysis of disparate and dynamic data streams. Horizontal scalability and linear throughput are supported in the heart of the platform itself. It is being used to support very broad application areas, ranging from high-resolution settlement mapping, national bioenergy infrastructure to urban information and mobility systems. The platform provides spatio-temporal decision support capabilities in planning, policy and operational missions for US federal agencies. Also, the platform is designed to be functionally and technologically sustainable for continued support of the US energy and environment mission for the coming decades.

R. Karthik (✉) • A. Sorokine • D.R. Patlolla • C. Liu • S.M. Gupte • B.L. Bhaduri
Geographic Information Science and Technology Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: karthikr@ornl.gov; sorokina@ornl.gov; patlolladr@ornl.gov; liuc@ornl.gov; guptesm@ornl.gov; bhaduribl@ornl.gov

P.-P. Mathieu, C. Aubrecht (eds.), *Earth Observation Open Science and Innovation*,
ISSI Scientific Report Series 15, https://doi.org/10.1007/978-3-319-65633-5_7

Introduction

In today's world, there is a multitude of heterogeneous Earth Observation High-performance Computing (EO-HPC) systems, each designed to solve specific science or technological missions. These EO-HPC systems often work independently of each other, hindering the flow of information and limiting the ability to achieve interoperability among systems. Bringing these systems together to create a fully and closely knitted Cyber-infrastructure (CI) platform provides a shared universe for EO data driven analysis and discovery capabilities for US federal agencies. Though there have been some promising community efforts to build a CI platform that can integrate various types of EO-HPC systems together, there does not exist a unified CI platform currently (Kalidindi 2015; Bhaduri et al. 2015a). With "Big Data" explosion changing the landscape of software architecture design, there is an increasing need for such a platform that can meet the modern complex needs.

Geographic Information Science and Technology (GIST) and The Urban Dynamics Institute (UDI) at Oak Ridge National Laboratory (ORNL) are building such a CI platform by integrating our next-generation EO-HPC systems under one umbrella using novel techniques and emerging technologies. We employ wide breadth of techniques across the spectrum of scientific computing such as GPGPU, distributed, real-time and cluster computing (Bhaduri et al. 2015a; Karthik 2014a; Sorokine et al. 2012). The platform is designed to scale to petabytes of data and handle massive workloads. One of the biggest architectural challenges in developing our CI platform was addressing complex interdependencies among the various systems without compromise in efficiency or functionalities of individual systems. The platform achieves foundational, structural and semantic interoperabilities to create a harmonized and seamless experience across various systems. With this platform, various systems are designed to work together as a whole information system, but also retain the ability to operate independently if desired. In this integrated platform, the systems communicate with each other providing various levels of control of components and functionality, yet allowing for independent control as well. Modularity is being designed at the core of the CI platform to support future EO-HPC systems for easier integration, and foster sustainable development to meet US science and energy missions now and into the future. Our CI platform aims to take the next major leap in Data Science and Cyber-infrastructure, while making the best use of our existing EO-HPC systems.

In the following sections, we describe the challenges, trends and how our CI platform plays an important role in our research initiatives illustrated with settlement mapping, mobility science, and urban information system.

Settlement Mapping Tool (SMTOOL)

Understanding high-resolution population distribution data is fundamental to reduce disaster risk, eliminate poverty, and foster sustainable development. Modern censuses fail to cover population in remote, inaccessible areas in many underdeveloped nations. Small settlements are visible in high-resolution satellite imagery, which have historically been computationally expensive for information extraction. Utilizing GPUs, Oak Ridge National Laboratory is mapping the smallest settlements and associated population across the globe for the first time in our history. The high-resolution settlement and resulting 90 m LandScan HD population data have profoundly enhanced our ability to reach and serve vast vulnerable populations from local to planet scales (Bhaduri et al. 2015b).

Past 50 years have witnessed the global population increase by four billion and with 150 new births every minute, an additional four billion people will settle on this planet in the coming 50. Urban and rural population distribution data are fundamental to prevent and reduce disaster risk, eliminate poverty and foster sustainable development. Commonly available population data, collected through modern censuses, do not capture this high-resolution population distribution and dynamics. However, there is gross underestimation of global human settlements and population distribution. Footprints of our expanding activities are impacting the future of this planet from availability of natural resources to a changing climate. Accurate assessment of high-resolution population data is essential for successfully addressing key issues such as good governance, poverty reduction strategies, and prosperity in social, economic and environmental health. Geospatial data and models offer novel approaches to disaggregate Census data to finer spatial units; with land use and land cover (LULC) data being the primary driver. With increasing availability of LULC data from satellite remote sensing, “developed” pixels have been nucleus to assessing settlement build up from human activity. However, the processing and analysis of tera to peta scale satellite data has been computationally expensive and challenging.

With the availability of moderate to high resolution LULC data derived from NASA MODIS (250–500 m) or Landsat TM (30 m) have facilitated the development of population distribution data at a higher spatial resolution such as Oak Ridge National Laboratory’s (ORNL) LandScan Global (1 km) and LandScan USA (90 m); two finest resolution population distribution data developed. Although these LULC data sets have somewhat alleviated the difficulty for population distribution models, in order to assess the true magnitude and extent of the human footprint, it is critical to understand the distribution and relationships of the small and medium-sized human settlements. These structures remain mostly undetectable from medium resolution satellite derived LULC data. For humanitarian missions, the truly vulnerable, such as those living in refugee camps, informal settlements and slums need to be effectively and comprehensively captured in our global understanding. This is particularly true in suburban and rural areas, where the population is dispersed to a greater degree than in urban areas. Extracting settlement information from very high-resolution (1 m or finer), peta-scale earth observation imagery has been a promising pathway for rapid estimation and revision of settlement and population distribution data. As early as 2005, automated feature extraction algorithms implemented on available CPU-based architectures demonstrated radical improvement in image analysis efficiency when manual settlement identification from a 100-km^2 area was reduced from 10 h to 30 min. However, this scaled inefficiently with limited resources on a workstation and at that rate processing 57-million km^2 habitable area would take decades. The pressing need for identifying population distribution in the smallest human settlements and monitoring settlement patterns at local to planet scale as landscape changes are induced by population growth, migration, and disasters, compels a computational solution to process large volumes of very high resolution satellite imagery. Such a solution did not exist before this work was accomplished at ORNL.

Extracting settlements at high-resolution satellite images was accomplished by mapping sub-meter pixel data to unique patterns that correlate with the underlying settlements. To account for the variations in settlement structures spanning from skyscrapers to small dwellings, we generated patterns at different scales. The mapping process involved simultaneously analyzing a set of connected pixels to extract low-level structural features such as building edges, corners, and lines. Next, the spatial arrangements of these structural features at different scales were computed in parallel allowing us to generate unique settlement signatures efficiently. Furthermore, pattern recognition based on a previously learned model was integrated with the pattern generation step allowing us overcome the need to for additional storage. Our strategy of mapping pixels to underlying structural patterns was quite different from existing approaches that relied on spectral measurements such as reflectance at different wavelengths for settlement detection.

This approach using sub-meter resolution imagery is not only useful in generating accurate human settlement maps, but also it allows potential (social and vulnerability) characterization of population from settlement structures (tents, huts, buildings) from image texture and spectral features. Rapid ingestion and analysis of high resolution imagery to enhance quality and timely availability of input spatial data provides a cost and time effective solution for developing current and accurate high resolution population data. Such progresses in geospatial science and technology hold tremendous promise for advancing the state of accuracy and timely flow of critical geospatial information not only to benefit numerous sustainable development programs; but also has significant implications for time critical missions of disaster support.

High-resolution settlement data is foundational information for locating populations and activities in an area. One major usage of this data is as input to ORNL's LandScan HD population distribution model, which combines the settlement data with population density information to generate population distribution at an unprecedented 90 m resolution. For many underdeveloped countries, official censuses never reach remote, difficult to access areas. Moreover, there have not been reasons to exploit high-resolution satellite imagery for those areas. Consequently, this capability has enabled us to locate populations in secluded areas of the planet for the first time in our history. High-resolution settlement and population data are being used by the global humanitarian community for missions ranging from planning critical infrastructure and services to the deserving population, responding to the Ebola crisis in western Africa, eradicating Polio in Nigeria, as well as defining and mitigating disaster risks.

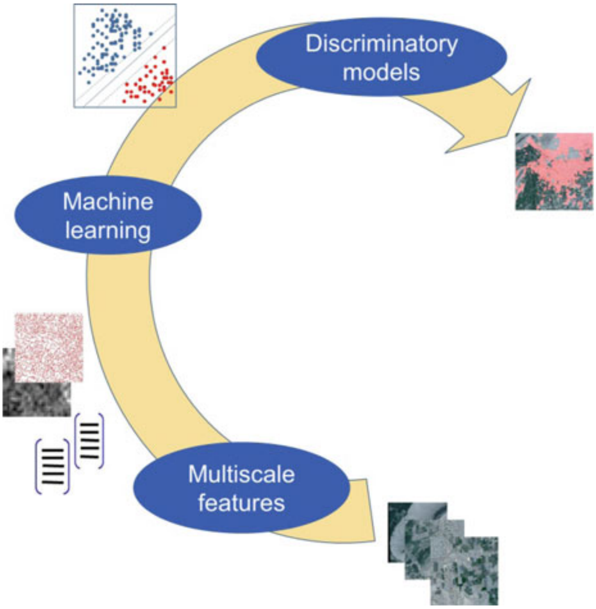


Fig. 1 An overview of SMTOOL architecture

Often solutions require advanced algorithms capable of extracting, representing, modeling, and interpreting scene features that characterize the spatial, structural and semantic attributes. Furthermore, these solutions should be scalable enabling analysis of big image datasets; at half-meter pixel resolution the earth's surface has roughly 600 Trillion pixels and the requirement to process at this scale at repeated intervals demands highly scalable solutions. Thus, we developed a GPU-based computational framework (as illustrated in Fig. 1) designed for identifying critical infrastructures from large-scale satellite or aerial imagery to assess vulnerable population. We exploit the parallel processing capability of GPUs to present GPU-friendly algorithms for robust and efficient detection of settlements from large-scale high-resolution satellite imagery (Patlolla et al. 2012). Feature descriptor generation is an expensive (computationally demanding), but a key step in automated scene analysis. To address the large-scale data processing needs we exploited the parallel computing architecture and carefully designed our algorithm to scale with hardware and fully utilize the memory bandwidth (required to transfer the high resolution image data) efficiently to produce great speedups times for the feature descriptor computation (as illustrated in Fig. 2).

We could thus achieve GPU-based high speed computation of multiple feature descriptors—multiscale Histogram of Oriented Gradients (HOG) (Patlolla et al. 2012), Gray Level Co-Occurrence Matrix (GLCM) Contrast, local pixel intensity statistics, Texture response (local Texton responses to a set of oriented filters at each pixel) (Patlolla et al. 2015), Dense Scale Invariant Feature Transform (DSIFT), Vegetation Indices (NDVI), Line Support Regions (extraction of straight line segments from an image by grouping spatially contiguous pixels with consistent

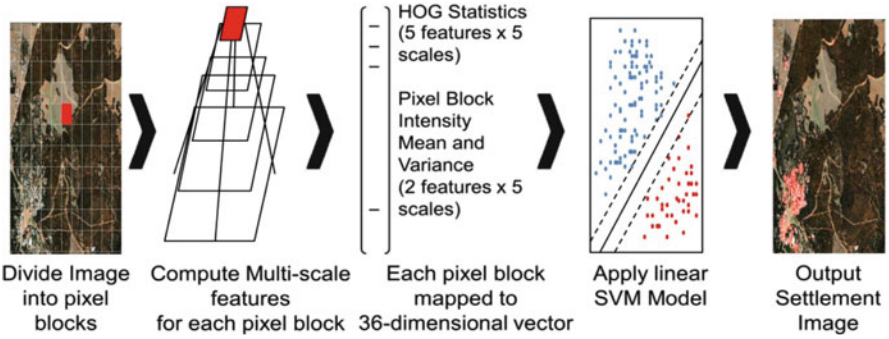

HOG Statistics (5 features x 5 scales)
Pixel Block Intensity Mean and Variance (2 features x 5 scales)
Divide Image into pixel blocks
Compute Multi-scale features for each pixel block
Each pixel block mapped to 36-dimensional vector
Apply linear SVM Model
Output Settlement Image

Fig. 2 Settlement mapping process

orientations), Band Ratios (a digital image-processing technique that enhances contrast between features by dividing a measure of reflectance for the pixels in one image band by the measure of reflectance for the pixels in the other image band) etc. Once, the features are computed, a linear SVM is used to classify settlement and non-settlements. The computational process requires dozens of floating point computations per pixel, which can result in slow runtime even for the fastest of CPUs. The slow speed of a CPU is a serious hindrance to productivity for time critical missions. Our GPU-accelerated computing solution provides an order of magnitude or more in performance by offloading compute-intensive portions of the application to the GPU, while the remainder of the code still runs on the CPU. The implementation further scales linearly with the available nodes thus enabling the processing of large-scale data on high end GPU-based cluster-computers.

With the introduction of very high-resolution satellite imagery, mapping of small or spectrally indistinct settlements became possible on a global scale. However, existing methodologies for extracting and characterizing settlements rely on manual image interpretation or involve computationally intensive object extraction and characterization algorithms that saturate the computational capabilities of conventional CPU-based options used by commercial remote sensing software packages. Many of these existing pixel based image analysis techniques used for medium resolution Landsat imagery (∼30 m) or coarser MODIS imagery (250–1000 m) are not ideal for interpreting satellite imagery with sub-meter spatial resolution. Advanced modeling of the spatial context is necessary to extract and represent information from such high-resolution overhead imagery. On a global scale, critical computational challenges are posed in the processing of petabytes of sub-meter resolution. For example an image of Kano city in Nigeria at 0.5 m resolution represents 23 Gigabytes of data covering 13,050 km^2. Attempts to accurately extract settlements using CPU-based commercial remote sensing packages were unsuccessful. An estimate to manually digitize the settlements for this area was 870 h. Meanwhile, SMTool on a 4 Tesla GPU workstation is able to process this large dataset in approximately 17 min (as illustrated in Table 1, Figs. 3 and 4).

Table 1 Performance of various features—processing times are based on a 4 C2075 GPU workstation

Feature	Accuracy (%)	Runtime (s)
HOG	93.5	1.6
TEXTONS	92.7	4.7
VEGIND	91.4	1.77
BANDRT	86.1	1.93
DSIFT	90.8	11.33

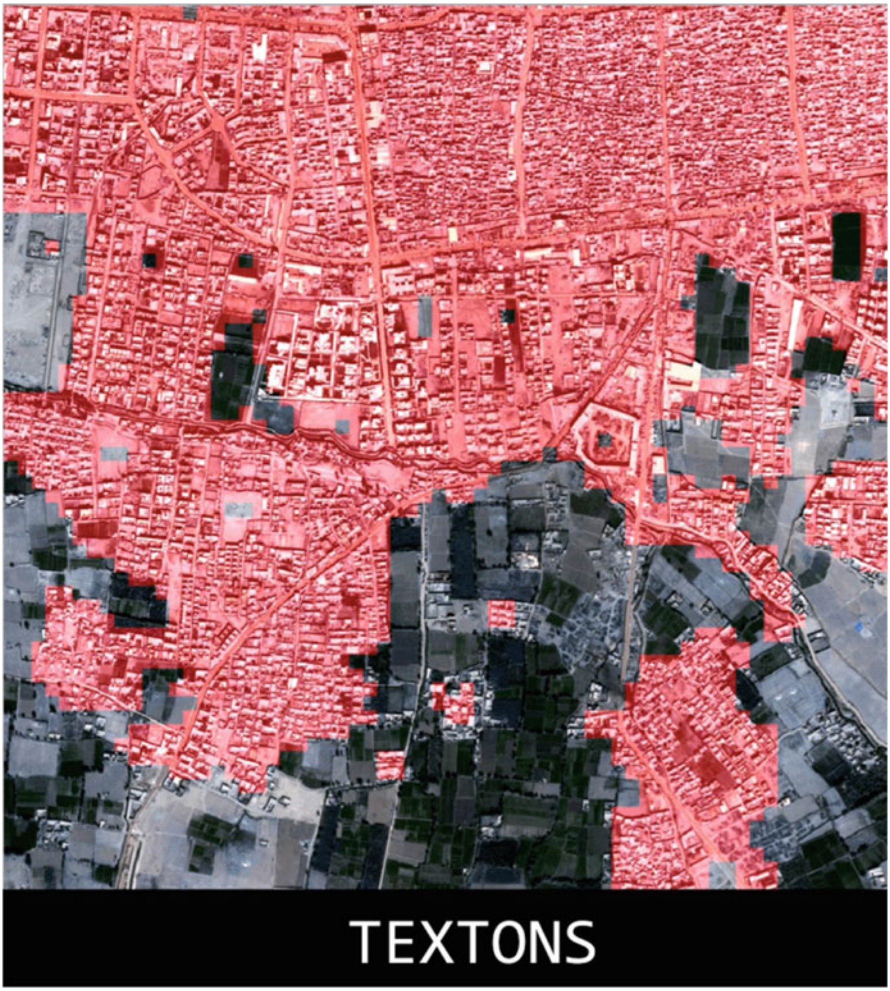


Fig. 3 TEXTONS performance—33 images of 0.5 m spatial resolution, each covering an area of 2.6 km^2, collected from various parts of Kandahar, Afghanistan

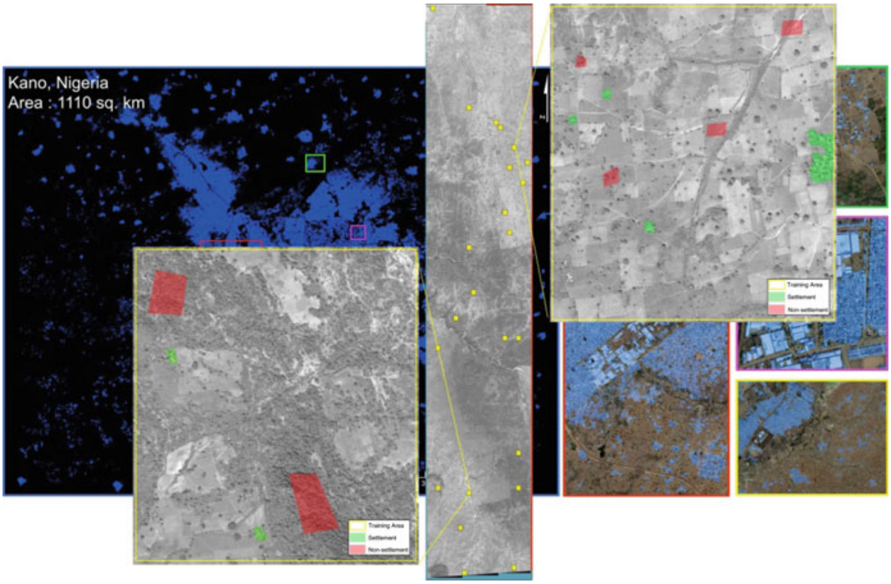


Fig. 4 SMTOOL results for Kano, Nigeria—On an average ∼1% training samples

The Compute Unified Device Architecture (CUDA) has enabled us to efficiently utilize NVIDIA TESLA GPUs to develop and utilize (in a practical and efficient manner) the expensive feature descriptor algorithms that would otherwise be complex and impractical.

It is quite important to carefully design the computing strategies to fully exploit GPU's parallel computational capabilities. For this we divide the high-res imagery into non-overlapping square pixel-blocks consisting of $M \times M$ pixels. To set an optimal value for M, we experimented with several values ranging from 8 to 64 pixels. For each of the pixel-block we compute feature descriptors at several scales based on the low-level image features. Parallel processing is key to efficiently implementing the feature descriptor algorithm, where thousands of threads can run simultaneously. The K20X Tesla GPU in the GEOCLOUD cluster has 2688 CUDA cores/GPU at ∼3.95 Tera Flops of single precision computing power, which helped us deliver speedups ranging from 100× to 220×, thus cutting the feature descriptor computation time from days to mere minutes.

Efficient memory utilization is key for the optimal performance of our algorithm on the GPU as it involves transfer of large amounts of image raster data to the GPU and the output settlement layer back to the host. We leveraged the Tesla GPUs PCI-E 3.0 interface to achieve over 10 GB/s transfers between the host (CPUs) and the devices (GPUs). An important factor to consider is the data transfer between CPU and GPU and optimal speedups require reducing the amount of data transferred between the two architectures. In our implementation the data transfer between CPU and GPU is performed only at the beginning and the end of the process. First, the image is read using GDAL to store the data in the CUDA global memory, which is 6 GB and provides a 288 GB/s memory–memory bandwidth. Though the data requests to the global memory has higher latency, CUDA provides a number of additional methods for accessing memory (i.e. shared, constant memory etc.) that can remove the majority of data requests to the global memory, thus enabling us to keep the Compute to Global Memory Access (CGMA) ratio at high values to achieve fine grained parallelism.

This has been an interdisciplinary effort with team members with academic and professional training in geography, electrical engineering, computer science and engineering and expertise in geographical sciences, population and settlement geography, spatial modeling, satellite remote sensing, machine learning data analysis and high performance computing.

Toolbox for Urban Mobility Simulations (TUMS)

TUMS, Toolbox for Urban Mobility Simulations, is a web-based high-resolution quick response traffic simulation modeling system for urban transportation studies. TUMS can be used both as a daily commuter traffic simulator or an emergency evacuation planning tools. There are some unique features in TUMS comparing with other similar transportation modeling and traffic simulation systems. It uses high-resolution population distribution and detailed street network, both covering the entire world. TUMS is aiming to simulate county level traffic flow using microscopic traffic simulation modeling and it has web applications based on WebGL. Users can take advantage of client side Geographic Process Unit (GPU) when it presents on client desktop. The transportation engine of TUMS is based on an open source package called TRANSIMS (TRansportation ANalysis SIMulation System, version 5.0) (Smith et al. 1995).

Global Dataset

Two main datasets used in TUMS are a population distribution dataset called LandScan developed by ORNL (Bhaduri et al. 2002) and a worldwide open source street level transportation network, called OSM, OpenStreetMap (OpenStreetMap 2016). Both datasets covers the entire planet.

LandScan has two components, LandScanUSA and LandScanGlobal. As the name indicates, LandScanUSA is the population distribution for USA and LandScanGlobal covers the entire world including USA. Both dataset are updated yearly. LandScan divides the study area into cells and each cell has a population count. LandScanUSA has higher resolution cells than LandScanGlobal. LandScanUSA uses 3 arc second cells while LandScanGlobal uses 30 arc second cells. Roughly, the 3 arc second cell has the size of 90 m by 90 m and the 30 arc second cell has the size of 1 km by 1 km around the equator. The size of cells becomes smaller when the latitude is higher. In order to make the analysis consistent, TUMS decomposes LandScanGlobal to 3 arc second cells using a primitive moving average method. If the study area is within USA then TUMS uses LandScanUSA dataset. If the study area is outside of USA then TUMS uses the decomposed LandScanGlobal 3 arc second dataset. From now on in this paper, we will use LandScan to represent both LandScanUSA and decomposed LandScanGlobal with 3 arc second cells.

OSM is updated weekly. The data quality in OSM depends on geographic region. Europe and North America data has much higher quality than Asia and Africa. Since OSM keeps evolving, the data quality now has improved tremendous compared to earlier version. Please do not be conceived by its name, OpenStreetMap, it not only have street network, it has other features such as land use type, administration boundary, physical features and lots more. However, TUMS only uses street network at current stage.

Framework

There are three major components in TUMS framework, a pre-processing component, a traffic simulation component, and a web-based visualization component (Fig. 5). The pre-processing component is responsible for preparing the input data for the transportation modeling. This first step is to define a study area, which can be a county (in USA only), a polygon or a circle. The next step is to extract the population and street network from LandScan and OSM. After integrated these two

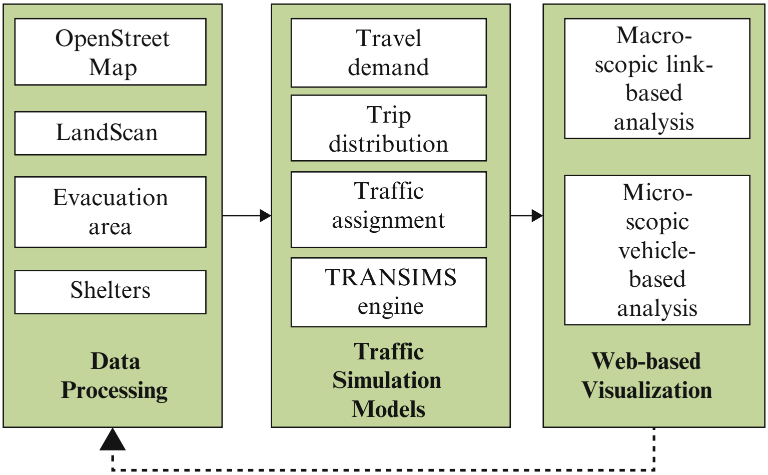
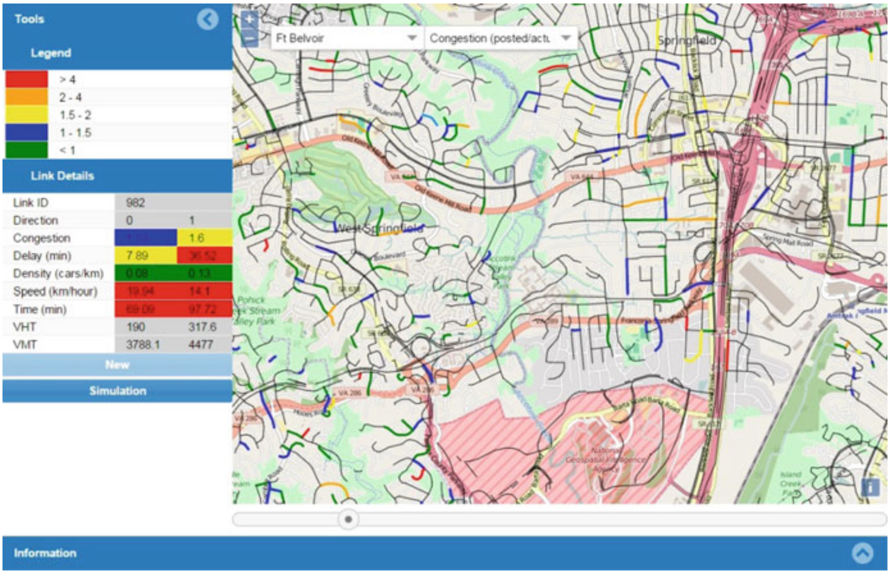


Fig. 5 TUMS framework

data together, TUMS creates a routable network with correct network topology and generates origin-destination (OD) tables for transportation modeling.

The traffic simulation models are based on TRANSIMS framework. TRANSIMS has more than a dozen executable programs, which are loosely coupled. Each executable program can be executed separately if the input data is properly set. Roughly, these programs can be grouped into five categories, synthetic population generation, network preparation, origin-destination (OD) table preparation, trip distribution and assignment, and Microscopic traffic simulation. TUMS takes the advantages of this flexible framework and integrate its own modules into TRANSIMS framework. For example, the synthetic population generation modules are replaced by LandScan population module. TUMS own OD table generation modules using LandScan and OSM substituted TRANSIMS OD table preparation modules.

Since TRANSIMS does not have a Graphic User Interface (GUI), TUMS developed two independent visualization tools for different background users (Karthik 2014b). The link based visualization and analysis tools are for planners who are interest in the measure of efficacy (MOE) for the planning purpose. The vehicle base animation tools are for traffic engineers who are more interested on operations such as intersection traffic control. Figures 6 and 7 are the examples for link-based and vehicle based GUI.



Link ID	982	
Direction	0	1
Congestion	[illegible]	1.6
Delay (min)	7.89	36.52
Density (cars/km)	0.08	0.13
Speed (km/hour)	19.94	14.1
Time (min)	69.09	97.72
VHT	190	317.6
VMT	3788.1	4477

Fig. 6 Link-based visualization tool

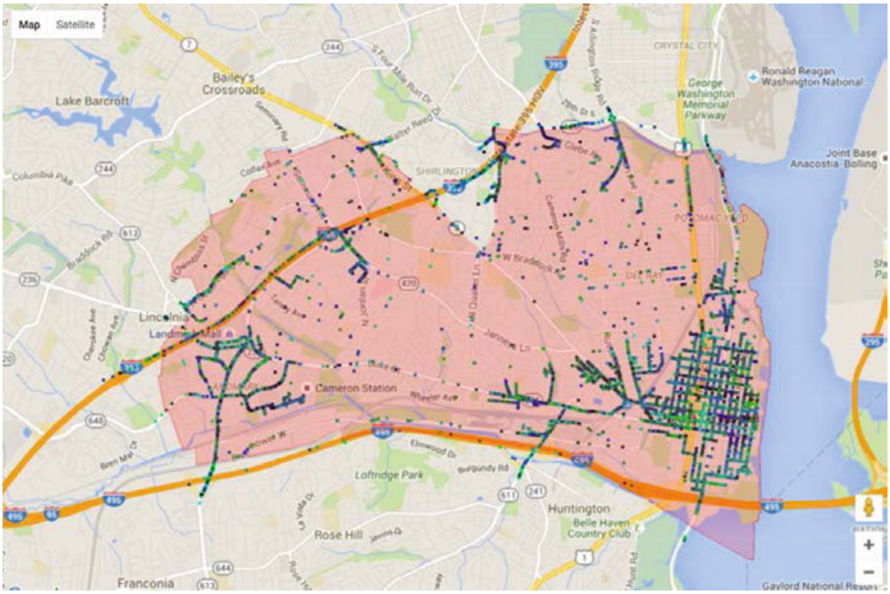


Fig. 7 Vehicle-based visualization tool

OD Tables

By manipulating the OD tables TUMS can simulate both daily commuter traffic flow or non-notice emergency evacuation simulation. Although LandScan only reports the total population count for each cell, but internally, LandScan has five layers, which are worker, residential, school, shopping and non-movement group. With these layers it is possible to generate the O-D tables for daily commuter traffic flow. Figure 8 is an example of daily commuter traffic flows for the year of 2015 and 2035.

For non-notice emergency evacuation, TUMS assumes that every evacuee would like to take a trip to the nearest shelter or exit point (boundary points) to get out of the evacuation area as quickly as possible. The OD tables for non-noticed emergency evacuation simulation are generated by finding the nearest shelters for each LPC.

Resolution

Traditional transportation models, both macroscopic and microscopic, use Traffic Analysis Zone (TAZ) for the OD tables. TAZs are the basic geographic unit for demographic data and land use type. The size of TAZ varies. Zones are smaller in urban area with high population density and larger in rural area with

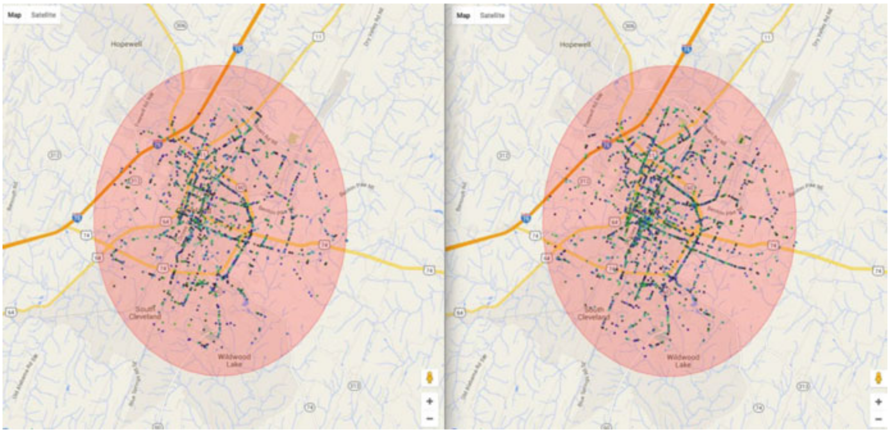


Fig. 8 Daily traffic flow simulation at Cleveland, TN, on the year 2015 (*Left*) and 2035 (*Right*)

lower population density. With the rising of agent-based and driver behaviour traffic modeling in transportation research, the large area TAZ is not suitable for microscopic traffic simulations. For example, Alexandria County, VA, has only 62 TAZs. Each TAZ covers quit large area. For microscopic traffic simulation there is no reason that the TAZs could not be as small as a single building if there is enough computing resource and the data available. The computing resource is cheap now, but unfortunately, the global single building population distribution database is not available yet. So TUMS uses LandScan as an alternative. There are 5657 LandScan Population Cells (LPC) comparing with 62 TAZs in Alexandria, VA. The LPC resolution is around 100 times higher than using TAZ.

The size of TAZ or LPC is related with the network level of details. A network with principle and minor arterials does not need high-resolution population dataset. In traditional traffic modeling the collect or local streets are ignored due to the low traffic volume. But if the OD zones use single buildings as the trip generation unit, then the network should include collect and local streets. For non-notice emergency evacuation simulation, the collect and local streets become very important because the evacuees who are close to the boundary of evacuation region can get out the evacuation area very quick by using local streets. If the local streets are excluded from the network, all these evacuees have to travel to the opposite direction in order to access the major arterials and then travel to the boundary points. This is unrealistic and generates artificial congestion on the arterials.

Unified Network and Population Database

There are many transportation modeling systems and traffic simulation tools available both on open source or in commercial packages, such as TRANSIMS, MITSIM, VISSIM, SUMO and MATSIM, just to name a few. All of these models have similar basic input requirements such as network and population. But each one of them has its own input and output format. TUMS has developed a uniform database for the entire world and also has utilities to convert the unified database to different format for different models. Currently TUMS supports TRANSIMS and MITSIM. SUMO and MATSIM will be added in the near future.

Big Data

In modern era like todays, Data is generated by internet activity, sensors for environment, traffic cameras, satellite imagery and is referred to as Big Data. Processing, analyzing and visualizing this data have its own challenges. The OSM dataset has 3+ billion point and 300+ million ways (links) in the planet data file (December 2015 version). Among the 300+ million ways there are 80+ million street links. The LandScan has 93+ billion cells.

Since all these data are read only, TUMS chooses flat binary files to store the data for its simplicity and for easy random access. In order to retrieve the data efficiently, both the network and the population data are decomposed to 1 by 1 degree cells. The street network is stored in a shapefile format and the population data is store in a binary grid format. Since ESRI's binary grid format is a proprietary format, TUMS has to develop its own binary grid format. In TUMS database the OSM street network occupies 21 GB and LandScan occupied 4.6 GB disk space.

The vehicle trajectory data is another challenge. Assume that there are 100K vehicles in a median size county and the simulation time is 24 h for a daily commuter traffic simulation with 1-s simulation step, the total trajectory for all vehicles is 8+ billion points per scenario. This data is stored in ESRI-like point shapefile. Since ESRI's point shapefile has 2 GB file boundary limitation, TUMS developed its own binary format without the 2 GB boundary limitation. This vehicle trajectory data is streaming to TUMS web-based application for animation.

Urban Information System (UrbIS)

Leveraging Big Data to Understand Urban Impact on Environment and Climate

Cities are one of the major contributors to climate change. At the same time cities themselves are most strongly affected by the changing environmental conditions. Immense complexity of interactions between urban areas, climate, and natural environments presents scientists with a multitude of challenges. First, urban environments are characterized by a very large number of variables including demographics, energy, quality of the environment and many others. Second, cities show significant variety and strongly differ in terms of their processes and energy and material flux. Third, cities themselves have become major defining forces for their surrounding environments by affecting local topography, air circulation, water-heat balance and habitats. Resulting human-natural system has a large number of feedback loops and correlations among its variables.

Understanding of the urban environments can be improved by tapping into vast information resources that have become available to the researchers thanks to the Big Data technologies (Chowdhury et al. 2015). Traditional sources of Big Data like historical databases of Twitter messages, postings in other social networks, and cell phone locations can provide valuable insights into the functioning of people in urban environments. However, the majority of the data of interest for urban researchers exist in the form of an "ecosystem of small data", as a large number of disparate datasets created by different communities, government agencies, and research institutions. Finding such data and then merging them together for the use in a single analytical workflow has become a major hindrance for such studies. To address these challenges we at ORNL have embarked on the developing of ORNL Urban Information System (UrbIS)—a web-based software tool that would allow urban scientists to perform most of their analytical and data processing tasks in the cloud within a unified browser-based user interface.

UrbIS goal is to address a number of problems typically faced by the researchers in this area. After analyzing ORNL experience in a number projects including the ones described in this chapter we were able to identify multiple bottlenecks that impede scientists' productivity. These challenges can be mitigated by developing software for automating of several commonly performed tasks such as (1) finding the data necessary to achieve the goals of the study, (2) preparing the data from external sources for the use in the analytical software, (3) running modeling and analytical programs on the high-performance computing systems, and (4) retrieving and understanding the results of the analysis including representation of the results in visual form as graphs and maps.

Although scientists typically have a good understanding of the kinds of data they need for their research, finding specific datasets and not missing the relevant ones may be hard and time consuming. Most of the relevant data resides in the "deep web", i.e. not visible or not suitably indexed by general-purpose search engines like Google or DuckDuckGo. Therefore, such search engines often produce noisy results that require lots of manual filtering and verification or miss relevant data.

Search through dataset metadata provides a better alternative for finding scientific data. In the recent decade metadata has become a universally used tool for documenting large amount of data especially produced by the governmental, international, and other major research organizations. Multiple standardization efforts have generated several specifications that cover lots of aspects of important domain-specific knowledge necessary to precisely represents information about the data. Metadata search capabilities are currently available in many data archives and repositories such as, for example, NASA's Data Portal (https://data.nasa.gov/)

and DataONE Earth Observation Network (https://www.dataone.org/) supported by National Science Foundation.

Metadata search in most cases is more effectives than the use of general-purpose search engines because the metadata is structured and curated according to well defined standards. Users can filter through the data not only by the keywords or commonly used phrase but also by specific spatial, temporal or attribute information. For example, it is possible to limit the search by a specific sensor, variable, target area, time interval, or range of values. Certain results can be excluded from the search by using negation criteria that is not easily achievable in general-purpose search engines. However, typical metadata search requires interaction with multiple metadata search systems and familiarity with a variety of user interfaces and APIs.

After the necessary data have been identified the users have to extract relevant subsets of data (i.e. clipping a region of interest and/or limiting the data to a specific time interval) and move the data to their workstations. Sometimes the volume of the data can be very large like in the case of ensembles of global circulation models and can reach the volumes on the order of terabytes. Present-day hard drive costs are low enough not to be a limiting factor for storing data still movement of the large volumes of data over the network requires lots of time and special software like Globus Toolkit GridFTP[1]. The bigger problem is the maintenance of the harvested data on the workstation or local network storage that requires not only cataloguing of the data but also checking the dataset integrity, creating backups and retrieving updates for corrected errors or newer versions of the datasets.

The next preparation step is converting harvested data into the formats that can be understood by the analytical software. This step includes not only simple format conversion but also other non-analytical operations. Almost all of the urban data is spatiotemporal as the overwhelming majority of data records in urban datasets have some kind of geographic and time reference. Thus there is always a need to maintain and convert cartographic projection and other spatial referencing information. The datasets often come in the formats that are not understood by the analytical software or in-house developed code and scientists are forced to spend their time on developing format converters or perform lots of manual transformations. In case of UrbIS we are often faced with the data that comes from different scientific communities—urban scientists and climate modellers. Most climate and weather data is stored in NetCDF or HDF5 files while urban datasets mostly rely on the file formats of the commercial GIS software. Many of the open-source and commercial GIS are able to read these formats but the data have to be manually reorganized. The separate problem is semantics misalignment among the datasets especially when the datasets originate from different communities. This includes incompatibilities related to the units of measure, variable names, inconsistent naming of the grids and spatial regions. Such differences between the datasets are often not reflected in their metadata.

[1] http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/.

Many of the analytical and modeling projects at ORNL including the ones in this review heavily rely on high-performance computing systems and facilities. This includes conventional computing clusters, cloud-based systems and leadership massively parallel facilities like TITAN and Eos[2]. Developing, porting and using scientific code and managing applications and data on such systems require special technical skills and experience that are not commonly available.

Finally, the output of the high-performance models has to be presented in the form suitable for understanding and presentation. In our research domain this almost always means visualization in the geographic context with the help of advanced visualization tools found in the geographic information systems. At this point the modeling and analytical results should not only be converted into the formats understandable by GIS but also aligned with other pertinent geographic data.

Even though most of the outlined difficulties are technical in nature, they impose a significant toll on the scientists' time and increase overall costs of research. Moving these burdens from the scientist is one of the main goals of UrbIS. Earlier ORNL experience with similar systems has demonstrated efficiency of such approach. In 2010 ORNL has developed iGlobe—a desktop application for the geographic analysis of climate simulation data that combined server-side analysis and management of data with geographic visualization in a single workflow (Chandola et al. 2011). iGlobe is built around NASA WordWind Java[3] and allows users to retrieve the data from the data portals, process them on the server, and visualize analysis results on the desktop. Control of the server-side processing of the data is performed through desktop GUI using secure shell connection. Results of the analysis are presented using NASA WorldWind visualization component as interactive 2D or 3D geographic displays. With the advance of web, cloud and high-performance computing technologies we are leveraging our iGlobe experience at the new level of web-centric and cloud-centric applications.

Currently UrbIS is under active development and exists as an early evolving prototype available to ORNL internal users. When completed UrbIS will allow urban researchers to execute a complete analytical workflow starting with discovering and obtaining necessary data from diverse data repositories, analyzing them using high-performance computing capabilities, and then to visualizing and publishing the results. Researchers will be able to perform all these operations completely in the cloud and/or on the server through a standardized web interface. When fully implemented UrbIS will eliminate the need to download and process any input, intermediate, or output data files on the workstation.

Screenshots of the UrbIS prototype are presented from Figs. 9, 10, 11 and 12. UrbIS workflow starts with a federated metadata search interface (Fig. 9). This interface provides a user with the search capabilities through several external metadata search engines and internal ORNL data holdings. For the federated

[2] https://www.olcf.ornl.gov/titan/.

[3] http://worldwind.arc.nasa.gov/java/.

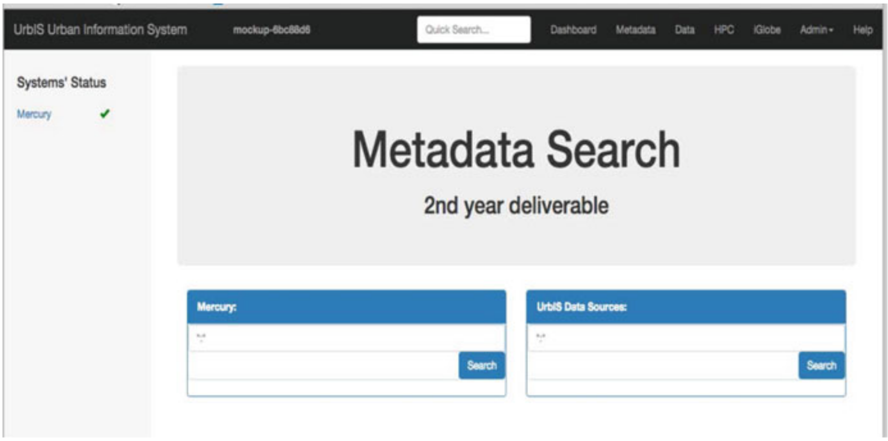
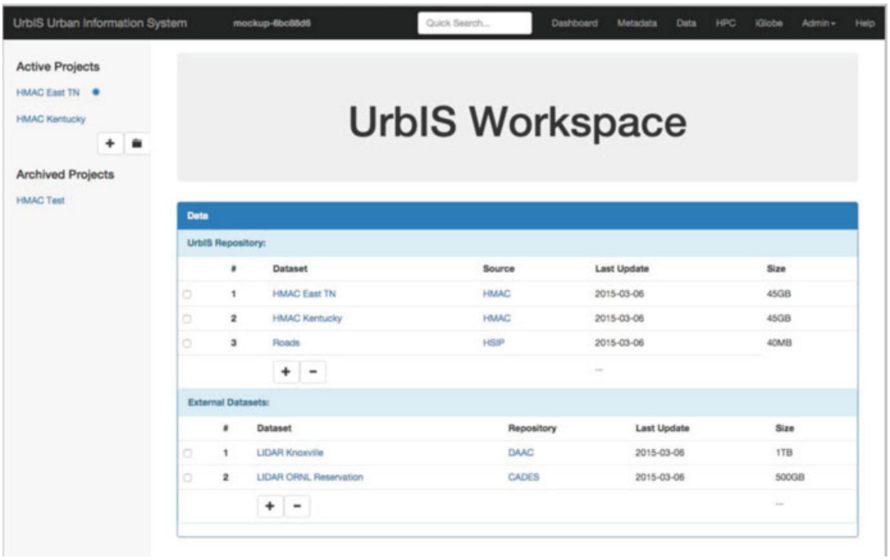


Fig. 9 UrbIS federated metadata search interface



UrbIS Repository:

	#	Dataset	Source	Last Update	Size
□	1	HMAC East TN	HMAC	2015-03-06	45GB
□	2	HMAC Kentucky	HMAC	2015-03-06	45GB
□	3	Roads	HSIP	2015-03-06	40MB

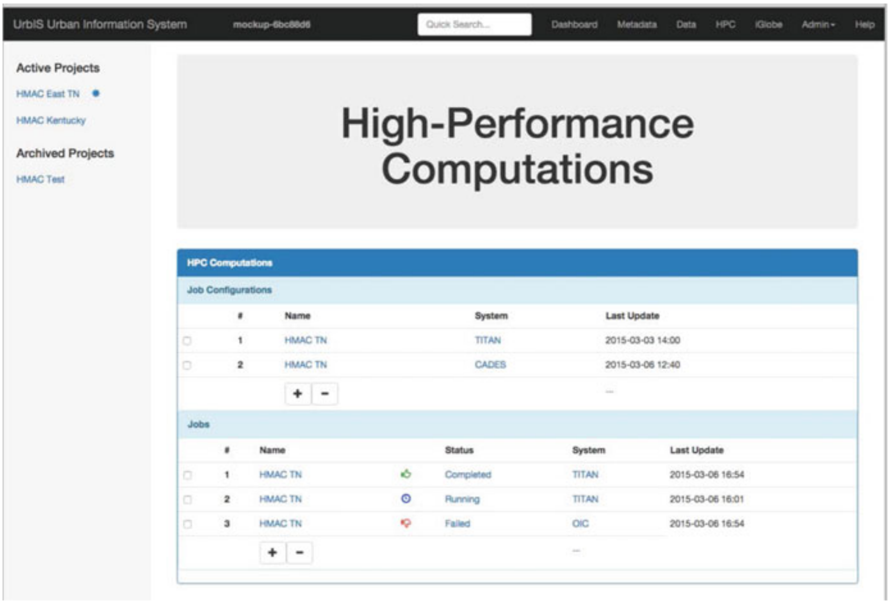
External Datasets:

	#	Dataset	Repository	Last Update	Size
□	1	LIDAR Knoxville	DAAC	2015-03-06	1TB
□	2	LIDAR ORNL Reservation	CADES	2015-03-06	500GB

Fig. 10 UrbIS workspace manager interface

metadata search engine we are using a customized version of Mercury[4]—an in-house ORNL metadata search engine that enables the search over other metadata repositories and archives like DataONE (https://www.dataone.org/) and ORNL DAAC (https://daac.ornl.gov/). In addition to the external data repositories UrbIS also provides its users with several frequently used datasets with common used

[4] http://mercury.ornl.gov/.



Active Projects

HMAC East TN ●

HMAC Kentucky

Archived Projects

HMAC Test

High-Performance Computations

HPC Computations

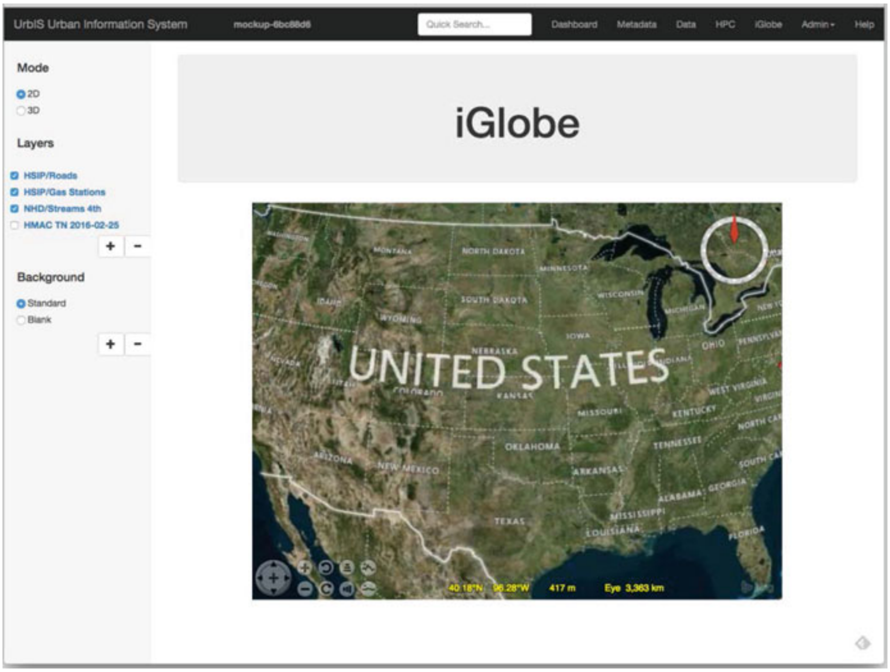
Job Configurations

	#	Name	System	Last Update
□	1	HMAC TN	TITAN	2015-03-03 14:00
□	2	HMAC TN	CADES	2015-03-06 12:40

Jobs

	#	Name		Status	System	Last Update
□	1	HMAC TN	👍	Completed	TITAN	2015-03-06 16:54
□	2	HMAC TN	◷	Running	TITAN	2015-03-06 16:01
□	3	HMAC TN	👎	Failed	OIC	2015-03-06 16:54

Fig. 11 HPC computations for analysis and modeling



iGlobe

Mode

● 2D
○ 3D

Layers

☑ HSIP/Roads
☑ HSIP/Gas Stations
☑ NHD/Streams 4th
□ HMAC TN 2016-02-25

Background

● Standard
○ Blank

Fig. 12 UrbIS visualization interface

geographic information. All the datasets and datastores can be searched through the same interface completely transparent for the user.

After finding the needed data the user defines the region of interest and spatial resolution of his study area. At the same time in the background the system starts retrieval and sub-setting of the requested data from the external data stores. After the data has been placed into UrbIS scratch disk space the user will be notified and the records about downloaded subsets will appear in the workspace manager interface (Fig. 10). Here users can check the statistics of the downloaded data and verify completion of the download and conversion processes. All the retrieved data will be stored cloud-side and will not be downloaded to the user's workstation unless requested. Internally the data will be converted into application-specific representations optimized for further processing and access through UrbIS web services.

At the next stage of the workflow the user will be able to choose from a library of the analytical and modeling functionality (Fig. 11). As a part of the initial UrbIS development we are implementing high-performance clustering algorithms for building typologies of the cities based on a large number of input parameters. After specifying input parameters the user will submit a task to one of the high-performance computers. UrbIS will prepare the data in the form suitable for the selected processing method and create a batch configuration file containing commands for the target high-performance platform. The user will be able to initiate processing on the target system directly from UrbIS interface. After the job completion UrbIS will retrieve the results and convert them into the formats that are used internally.

The final step of the workflow is the visualization of the results in the geographic context. For that purpose we are using WebWorldWind (https://webworldwind.org/)—a modern javascript version of NASA WorldWind that utilizes WebGL. It can be launched from within popular browsers without the need to download any plugins or desktop applications. Visualization section of UrbIS (Fig. 12) has user interface typical for a digital globe like Google Earth or NASA WorldWind. Here the user can visualize the input and output data in the geographic context. The data is fed to the visualization component with WMS and WFS services from the internal UrbIS storage. Also the user can pull the data from any other data source supported by the NASA WebWorldWind including default WorldWind layers. The user will have an ability to switch between 3D and 2D views and choose the background and portrayal methods most suitable for his visualization purposes.

Current implementation of UrbIS is being developed using nodejs for the server side components. As a spatial data storage we are using PostgreSQL with PostGIS extensions. High-performance processing components are implemented as external modules and they use languages and tools most appropriate for the specific algorithms and platform. UrbIS should be accessible from any modern browser with WebGL support enabled (for visualization component). Internally UrbIS relies on service-oriented architecture with most functionality exposed through RESTful programming interface.

Currently UrbIS is in the active development and is available for testing to internal users. Its implementation will enable users to use high-performance and cloud-based infrastructure in their research and reduce the time needed for mundane tasks such data movement and format conversion. Also UrbIS will serve as a testing ground for new cloud-based technologies to facilitate the use of large geodata in scientific research within high-performance and cloud-based environment. After initial release and testing with internal user community we will proceed to implementing other sets of functionalities and extend the library of the high-performance analytical routines with other methods and models. In the future we plan to integrate UrbIS infrastructure with systems like Jupyter Notebooks (http://jupyter.org/) so that users can develop their own code through a web interface and access UrbIS data using web services.

Conclusions

Efforts to understand and analyze data-enabled science has created a clear need to unite various Earth Observation High-performance Computing (EO-HPC) systems, where the best of these various worlds are brought together in one shared Cyber-Infrastructure (CI) platform. In this chapter, we have discussed such a CI platform being developed at Oak Ridge National Laboratory using data-driven GeoComputation, novel analytical algorithms and emerging technologies. Systems interoperability, scalability and sustainability play an ever-increasing role in data-driven and informed decision-making process in our platform. We have discussed architectural and technical challenges in development of our platform, and broadening implications of it as illustrated by our research initiatives for data and science production. With technological roots in HPC, our platform is optimized for Earth Observation Big Data used to accelerate the research efforts, and foster knowledge discovery and dissemination more quickly and efficiently for US federal agencies.

Acknowledgements The authors would like to thank a number of US federal agencies for their continued support for the research presented here. Sincere gratitude is due to many of our Geographic Information Science and Technology group colleagues for their collaboration and assistance. This paper has been authored by employees the US Federal Government and of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy. Accordingly, the US Government retains and the publisher, by accepting the article for publication, acknowledges that the US Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US Government purposes.

References

Bhaduri B et al. (2002) LandScan. Geoinformatics 5(2):34–37

Bhaduri B et al. (2015a) Emerging trends in monitoring landscapes and energy infrastructures with big spatial data. SIGSPATIAL Spec 6(3):35–45

Bhaduri BL et al. (2015b) Monitoring landscape dynamics at global scale: emerging computational trends and successes. Oak Ridge National Laboratory, Oak Ridge, TN

Chandola V et al. (2011) iGlobe: an interactive visualization and analysis framework for geospatial data. Proceedings of the 2nd International Conference on Computing for Geospatial Research & Applications, 23 May 2011, p 21

Chowdhury P et al. (2015) An comparison of data storage technologies for remote sensing cyber-infrastructures. The International Conference on Big Data Analysis and Data Mining

Kalidindi SR (2015) Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. Int Mater Rev 60(3):150–168

Karthik R (2014a) SAME4hpc: a promising approach in building a scalable and mobile environment for high-performance computing. Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, 4 November 2014, pp 68–71

Karthik R (2014b) Scaling an urban emergency evacuation framework: challenges and practices. Workshop on Big Data and Urban Informatics

OpenStreetMap (2016) https://www.openstreetmap.org. Accessed May 20 2016

Patlolla DR et al. (2012) Accelerating satellite image based large-scale settlement detection with GPU. Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data, 6 November 2012, pp 43–51

Patlolla D et al. (2015) GPU accelerated textons and dense sift features for human settlement detection from high-resolution satellite imagery

Smith L et al. (1995) TRANSIMS: transportation analysis and simulation system. Los Alamos National Laboratory, New Mexico

Sorokine A et al. (2012) Tackling BigData: strategies for parallelizing and porting geoprocessing algorithms to high-performance computational environments. GIScience