Large Scale Read Classification for Next Generation Sequencing

James M. Hogan¹* and Timothy Peut¹

¹: School of EECS, Queensland University of Technology, Brisbane, Australia.
j.hogan@qut.edu.au, tim@timpeut.com.

Abstract
Next Generation Sequencing (NGS) has revolutionised molecular biology, resulting in an explosion of data sets and an increasing role in clinical practice. Such applications necessarily require rapid identification of the organism as a prelude to annotation and further analysis. NGS data consist of a substantial number of short sequence reads, given context through downstream assembly and annotation, a process requiring reads consistent with the assumed species or species group. Highly accurate results have been obtained for restricted sets using SVM classifiers, but such methods are difficult to parallelise and success depends on careful attention to feature selection. This work examines the problem at very large scale, using a mix of synthetic and real data with a view to determining the overall structure of the problem and the effectiveness of parallel ensembles of simpler classifiers (principally random forests) in addressing the challenges of large scale genomics.

Keywords: Genomics, Next Generation Sequencing, Alignment-free Methods, Machine Learning

1 Introduction

Over the past decade, successive developments in sequencing technologies have revolutionized molecular biology, increasing sequencing efficiency as much as 100,000-fold over conventional Sanger sequencing, the foundation of the early public genome projects. These technologies, known collectively as Next Generation Sequencing (NGS) approaches [1], have fundamentally changed the nature of molecular biology, resulting in an explosion in the availability of genomic data sets, supporting routine sequencing in individual laboratories and, increasingly, as part of clinical practice.

The present work has its origins in clinical sequencing, and the need to identify rapidly the species present from the reads obtained from an NGS sequencing run. These ideas were considered in the context of bacterial infection, although the approach is more broadly applicable. Briefly, as described in [2], a lab may isolate DNA from a bacterial colony, and sequence the genome before the species is known with certainty. Clinical signs and non-molecular diagnostics may offer some insight, but the

* Author to whom all correspondence should be addressed.
suspicion needs to be confirmed if species-specific downstream informatics are to work successfully. The task is to determine, based on the FASTQ sequence files, whether the whole project is based on an homogeneous sequence sample, one reflecting a single species. In the alternative, the project might be corrupted, drawn from a mix of species, or reflect a previously unknown species or strain. Other confounding factors are outlined in the same reference.

The representation adopted in [2], tokenizing each read according to substrings of length k (kmers) and representing each project as a high dimensional count vector, proved extremely successful as a basis for SVM based discrimination of projects across both divergent and closely related species, with accuracy, precision and recall all well above 90%. Methods based on kmer spectra allow consideration of similarity without the need for assembly or alignment, and are robust in the presence of structural re-arrangement (see for example [3]). However, these results rely on careful and computationally intensive selection from the available set, a set whose dimension grows as \(4^k\).

While the earlier study was very successful as a machine learning exercise, and had considerable utility for the set of species considered, the more general problem has not been given significant attention. In particular, the effects of scale and sequencing noise, and computational methods to overcome them have not been previously investigated. The data sets in [2] were limited to around 130 projects, and the choices of \(k\) to 6, 8 and 10, with putative feature sets thus ranging from 4096 to 1048576. While feature selection limits the effective dimensionality, SVM training remains a non-trivial computational burden, and one which grows markedly with scale. SVM classification is usually parallellisable, but attempts to parallelise the training process have proven far less successful, the constrained quadratic programming problem admitting few obvious decompositions.

In this work we extend the previous study through direct consideration of the scale of the data set through a mix of data: real (drawn from the Sequence Read Archive [4]) and synthetic (generated by shattering completed and assembled genomes from GenBank through the use of the NCBI ART toolset [5]). Given the relative complexity and sequential structure of SVM training, we seek to use methods that may be more effectively used in systems which partition the training and classification problems to take advantage of the available computational resources. We consider primarily the environment provided by the open source Mahout Project [6] and the multiple classifiers provided through the Naive Bayes and Random Forests libraries.

This paper is organized as follows. In section 2, we provide more information on the nature of the problem and on the data sets utilized in the earlier study, before successively introducing the machine learning environment Mahout and the open source map reduce framework Hadoop which underpins it (section 2.1) and finally the Random Forest classification method used over the remainder of the paper (section 2.2). Section 3 is concerned with the data selection and preparation, and our focus is mainly on the use of ART and additional random substitutions to generate variation in the data set, before considering the computational demands of data preparation in section 3.1. Section 4 presents the results of our classification experiments and we conclude in section 5 with discussion and future extensions.

2 Background

Next Generation Sequencing (NGS) [1] is the name given to a number of DNA sequencing technologies, all of which build on classical shotgun sequencing. Laboratory sequencing machines randomly break DNA up into numerous short, independent segments, known as reads. The resulting data are stored in the industry standard FASTQ format.

The task is to classify whether the entire sequencing project belongs to a particular organism, here chosen to be the important pathogen *Staphylococcus aureus*. The data used by Hogan et al [2] consist of 130 unique sequencing projects. These projects include a number of *S. aureus* projects, other Staphylococcus strains, and other common bacteria such as *E. coli*, *Streptococcus* and *Chlamydia*. All
of the sequencing projects were for prokaryotes, specifically bacteria. These were actual sequencing projects, such that there was variation between different sequences of the same bacterium strain as well as error introduced in the sequencing process. These projects were manually selected and obtained from the Sequence Read Archive (SRA, formerly known as the Short Read Archive) [4].

As briefly described above, Hogan et al. [2] used a k-mer representation as the basis of this classification. A sliding window of length k was used to count the occurrences of each k-mer across each read. These counts were then accumulated over all the reads in the project and then normalised, resulting in a feature vector of substantial length for each project. For k=8 and k=10, these feature vectors were reduced using the Relief [7] feature selection algorithm, as not all features were significant in classification.

This earlier study used a small but broad set of 130 sequencing projects to train and test a Support Vector Machine (SVM) in the R statistical environment. Two classes were used for classification - STAPH_AUREUS and NOT_STAPH_AUREUS. Precision and recall of above 0.95 was achieved, and the classifier was able readily to distinguish between different species of Staphylococcus. As noted above, concerns over the scalability of this approach [8] led us to consider a collection of simpler – or at least more simply trainable – classifier elements to allow parallelization. These techniques are considered in the next section.

2.1 Map-Reduce, Hadoop and Mahout

MapReduce [9] is a distributed computing model adapted from the map-reduce pattern of functional programming. Originally developed in its present form by Google, the method is designed to scale with increased data sets and additional hardware, splitting the computation over a set of equivalent nodes in such a way that the results may be readily assembled into groups and collapsed into an atomic value. The Apache Mahout project [10] is an environment for machine learning which exploits the map reduce framework provided by the Apache Hadoop [11] project. Numerous tutorials exist at the Hadoop project site and elsewhere describing the structure of Map-Reduce problems and the details of the architecture. Specific issues encountered in a class of genomics problems are considered in some detail in [12]. The crucial issue in the present application is the ‘map-reducability’ of the machine learning method under consideration; numerous methods admit a map-reduce representation in the classification phase, but far fewer can be parallelized in this way in the training phase.

Mahout exists to produce distributed or scalable machine learning libraries. It builds on top of Hadoop to provide MapReduce based implementations of common classification, clustering and data mining algorithms. Mahout is a work in progress and only has a small number of classification methods implemented, currently: Logistic Regression, Naïve Bayes, Random Forest, Online Passive Aggressive, and Hidden Markov Models. Attempts have been made to implement an SVM in Mahout but no working implementations currently exist. SVMs are difficult to parallelise and existing parallel SVM implementations are not able to be transformed into a MapReduceable form [8]. While there are theoretical MapReduceable SVM models [13], there are at present no working implementations.

In this work we consider only the Mahout Random Forest classifier in detail. Our study included extensive trials using R based implementations of Naïve Bayes, Random Forests and Regularised Random Forest algorithms over the 130 project data set from [2]. While we do not report these results in detail here, the Naïve Bayes classifier performed poorly relative to the others and to the earlier SVM methods. The Random Forest approach was considered promising, achieving results comparable to those of the SVM, and the method and these results are described in the next section.
2.2 Random Forest Classifiers

Random Forests are ensemble methods which use a number of independent decision trees for classification. Each tree uses a random selection of features to make a classification decision, and the Random Forest classifies by selecting the mode of all the decision tree outputs. By using a large number of decision trees, a Random Forest is able to overcome the problems associated with using a random selection of features to make a decision.

Decision Trees are well-established classification methods that recursively split a data set into smaller sets based on the result of a test defined at each branch in the tree [14]. Starting with the root node, decisions are made until a leaf node is reached, at which point the appropriate label can be applied to the data point. The problem in the training process is to determine which feature to use at each decision to partition the data points.

A Random Forest overcomes this semi-random feature selection by constructing a large number of decision trees, in the hundreds or even thousands. Each tree can be constructed independently, which makes Random Forest construction MapReduceable, with the reduction step incorporating an overall decision on the classification of the example data vector. The Mahout implementation chooses splits based on an information gain criterion, but does not regularize the approach to limit overfitting. In the trial studies using the R caret package, we found that RF with regularization outperformed even the SVM approach - albeit marginally and within the standard error - with precision of 0.95 vs. 0.94. The simple RF performance was notably weaker, but it was found that this performance could be made comparable to RRF through more aggressive feature selection, ultimately yielding results comparable to the SVM. The 130 projects from [2] were split into train and holdout sets consisting of 90 and 40 projects respectively. 10-fold cross-validation was used on the training set to calculate mean accuracy and standard deviation, while the holdout set was used to calculate holdout accuracy, precision and recall. Trials using Mahout and k=10 yielded optimal selection at around 1000 features, as may be seen in Table 1.

| Num Features | Mean Accuracy | Std Deviation | Holdout Accuracy | Precision | Recall |
|--------------|---------------|---------------|------------------|-----------|--------|
| 10,000       | 0.93          | 0.08          | 0.93             | 0.86      | 1.0    |
| 1,000        | 0.93          | 0.08          | 0.95             | 0.94      | 0.94   |
| 100          | 0.93          | 0.09          | 0.9              | 0.82      | 1.0    |

Table 1 - Mahout Random Forest with Varying Number of Features

3 Data Sets

In addition to the original 130 SRA sequencing projects from [2], a wide range of examples was here created using the NCBI ART tool set. ART is a set of command line tools that generate synthetic NGS reads from a full DNA sequence. ART generates these reads by simulating the sequencing process, and has built-in, vendor specific read error models to reproduce error in the sequencing process. Single-ended Illumina reads are supported; as this was the hardware used to sequence the projects from the earlier, this mode was selected for consistency.

ART's error models are derived from large sets of actual sequencing data and represent error in the sequencing process. ART is unable to reproduce the natural variation between organisms of the same species; it is only able to emulate sequencing error. For single-ended Illumina reads, the primary sequencing error is base substitution. That is, bases are sometimes incorrectly read and substituted for another base.
It is important to note that past a certain point increasing the read coverage does not improve the classification results. The coverage must be high enough such that the sequencing error introduced by ART can be normalised out when counting the k-mers. From experimentation it was found that coverage of 10 was sufficient. As the k-mer counts are normalised, only the relative frequencies of each k-mer are considered when performing classification. Once the coverage is high enough for sequencing error to be normalised out, further increases only increase the time taken to prepare the data set; they do not improve the classifier performance.

The reference sequences for 20 negative strains and one S. aureus were obtained from RefSeq, the NCBI reference sequence database [15]. With these reference sequences ART can be used to generate data sets of any size. In Hogan et al. [2], the 70 negative sequencing projects came from approximately 30 different strains, but it was not possible to obtain reference sequences for all of these 30. Reference sequences for the more uncommon and obscure strains did not exist in RefSeq. The reference sequences obtained include multiple Staphylococcus strains and other common bacteria such as E. coli, Streptococcus and Chlamydia. The full list of reference sequences is in the supplementary material at http://eprints.qut.edu.au/69837/.

As noted in the documentation and elsewhere, ART does not simulate variation between different organisms of the same strain; its error model purely deals with sequencing error. To introduce variation between organisms, bases were randomly substituted throughout a generated project before it was tokenised and k-mers were counted. This was accomplished by the following (pseudocode) list comprehension, which was run for each read in every synthetic project:

```python
[base if rand() >= p else rand_choice("ACGT") for base in read]
```

That is, there is a p probability that each base will be randomly substituted for another base. A being substituted for A is here a valid substitution; no switch is forced. Additional substitution was only performed on projects in the holdout set. Unless otherwise noted, for all runs ART generated reads of length 75 with a coverage of 10, and the top 1,000 features from Relief were used.

### 3.1 Computational Considerations

Data preparation was performed on a node in QUT's Big Data Lab. This machine had an Intel Xeon E5-2609 processor with 8 HyperThreads, each running at 1.2GHz. There was a total of 256GB of memory, with approximately 50GB free at any one time. All data was mounted on a network disk, and the node used 64bit Red Hat Linux as an Operating System.

As previously outlined, there are four main data preparation stages. Each stage was timed, with the results and time complexity of each stage shown in the table below. 1040 projects were created, with 360 in the training/test set and 680 in the holdout set. The k-mers were collected with k=10, the previously demonstrated optimal value, and Relief selected the top 1,000 features.

| Stage         | Wall Time Taken (H:M:S) | Time Complexity                                      |
|---------------|-------------------------|------------------------------------------------------|
| Create Reads  | 0:55:41                 | \(O\)\(\text{num\_projects} \times \text{read\_coverage}\) |
| Count k-mers  | 1:20:23                 | \(O\)\(\text{num\_projects} \times \text{read\_coverage}\) |
| Create CSVs   | 0:30:41                 | \(O\)\(\text{num\_projects} \times \text{total\_features}\) |
| Feature Selection | 3:14:14              | \(O\)\(\text{num\_projects} \times \text{total\_features}\) |
| Total         | 6:00:59                 | \(O\)\(\text{num\_projects} \times \text{read\_coverage} + \text{read\_coverage} \times \text{total\_features}\) |

**Table 1 - Computational Statistics, Data Preparation**
The overall time complexity of the data preparation stage is linear in the number of projects, the read coverage, and the total features. This linear time complexity is necessary to allow scaling to large data sets. Polynomial or otherwise super-linear complexity would present a large barrier to scaling. It is important to note also that the total features increases exponentially with \( k \) as \( \text{num}_\text{features} = 4^k \).

The first two stages of data preparation, read creation and \( k \)-mer counting, are embarrassingly simple to parallelise. Each synthetic sequencing project can be created independently, and the \( k \)-mers in each project can be counted independently. These two stages required approximately 18 hours of CPU time, and without this inherent parallelism the data preparation stage would be a significant barrier to scaling to larger data sets. As the data set grows, the preparation stage can be easily run on more powerful hardware. The limiting factor in this work was the use of the Relief feature selection algorithm, which while selected for its performance, was nonetheless a sequential implementation. A parallel version – based on subsetting – is likely necessary for extreme scale studies of this nature.

4 Results

The data preparation process described above was performed to create data sets of various sizes, primarily to test the computational effects of scaling to large data sets. There was no additional substitution performed: \( p=0 \). An approximate 70/30 split was used, as this achieved the best results in earlier work. 10-fold cross validation was employed on the train set to determine the mean accuracy and standard deviation of the Random Forest. The Mahout Random Forest randomly selected 31 features for consideration per node, and 500 trees were constructed. Mahout Random Forest was able to perfectly classify all projects at various data set sizes, albeit trivially as the noise level was zero. The timings for data preparation are provided above, and the training and classification timings were trivial in comparison, requiring around 1 minute on commodity hardware.

In the remainder of this section, we will consider the dual effects of noise and scale on classification performance. The number of projects is held to vary between the base of 130 as in [2] and a maximum of 1040 as described above. The essential task in each case remains to distinguish a chosen species from others in the set, where the difficulty arises from natural sequence diversity – although as we have seen, this effect is limited – and from increasing levels of noise. Overwhelmingly, we find that a random forest classifier produces very similar results for the problem regardless of the size of the data sets. In this respect, the only interest lies in the computational efficiency of the process, and this is considered in the previous section. While some limitations will inevitably arise at extremes, we have established that the approach can work very effectively for a very large collection of data, training and testing on 360 projects and evaluating performance on a hold out set of almost 700 projects. In the next sections we will examine the effects of noise and species class on the effectiveness of the approach.

4.1 Noise

With the classifier shown to work on a trivially simple problem, it is of interest to determine how much variation between organisms can be present before the classifier breaks down. This is accomplished by means of additional substitution, i.e. \( p>0 \). As above, 10-fold cross validation was employed on the train set to determine the mean accuracy and standard deviation of the Random Forest. The Mahout Random Forest randomly selected 31 features for consideration per node, and 500 trees were constructed. The results are summarised below.
Table 3 - Mahout Random Forest Results, Additional Substitution

| Substitution Rate | Mean Accuracy | Std Deviation | Holdout Accuracy | Precision | Recall |
|-------------------|--------------|---------------|-----------------|-----------|--------|
| 0.025             | 1.0          | 0             | 1.0             | 1.0       | 1.0    |
| 0.05              | 1.0          | 0             | 1.0             | 1.0       | 1.0    |
| 0.075             | 1.0          | 0             | 1.0             | 1.0       | 1.0    |
| 0.08              | 1.0          | 0             | 0.994           | 1.0       | 0.988  |
| 0.085             | 1.0          | 0             | 0.938           | 1.0       | 0.876  |
| 0.09              | 1.0          | 0             | 0.731           | 1.0       | 0.46   |
| 0.1               | 1.0          | 0             | 0.601           | 1.0       | 0.22   |
| 0.11              | 1.0          | 0             | 0.512           | 1.0       | 0.02   |
| 0.2               | 1.0          | 0             | 0.5             | 1.0       | 0      |

As these results show, the classifier starts to break down at 8% additional substitution. Its precision is consistently high, while recall quickly drops to unacceptable levels. That is, the classifier has no false positives and the false negative rate quickly increased past 8% substitution. Some insight into this problem may be gained by considering the most common features in the representations. We note that the feature counts for S. aureus, are predominantly higher than the counts for the other negative organisms, and we can highlight this variation by defining an envelope of the negative counts

\[ E_j = \max(N_{1j}, N_{2j}, N_{3j}, ..., N_{nj}) \]

where \( E_j \) is the \( j \)th feature count of the envelope and \( N_{ij} \) is the \( j \)th feature count of the \( i \)th negative organism.

![Figure 1 - Top 100 Features, Negative Envelope, p=0.00 and p=0.08. Note the disruption of separation with increasing noise levels.](image)
A graph of S. aureus, the negative envelope, and the mean of these values is shown on the previous page at left. As the noise level in the data set is increased, we can see a substantial disruption of this clean feature-based separation across the species set (right).

### 4.2 Other Species Targets

With the Random Forest able to successfully classify STAPH_AUREUS and NOT_STAPH_AUREUS classes, it is of interest to determine whether the k-mer representation and Mahout's Random Forest are capable of classifying other organisms. In this case, E_COLI and NOT_E_COLI were used as the positive and negative classes. As above, $k=10$ was used for k-mer counting and Relief was used to select the top 1,000 features. 10-fold cross validation was employed on the training/test set to determine the mean accuracy and standard deviation. 31 features were randomly selected for consideration per node, and 500 trees were constructed. The results are summarised below. As these results show, the classifier is able to successfully classify E. coli sequencing projects. Its precision is constantly high, while recall quickly drops to unacceptable levels above $p=0.06$. This behaviour was observed with S. aureus, but it is interesting to note that the transition occurs at a lower value of $p$, with classification based on a markedly different feature set.

| Substitution Rate | Mean Accuracy | Std Deviation | Holdout Accuracy | Precision | Recall |
|-------------------|---------------|--------------|------------------|-----------|--------|
| 0.05              | 1.0           | 0            | 1.0              | 1.0       | 1.0    |
| 0.06              | 1.0           | 0            | 1.0              | 1.0       | 1.0    |
| 0.065             | 1.0           | 0            | 0.9              | 1.0       | 0.81   |
| 0.07              | 1.0           | 0            | 0.67             | 1.0       | 0.34   |
| 0.075             | 1.0           | 0            | 0.55             | 1.0       | 0.103  |
| 0.08              | 1.0           | 0            | 0.51             | 1.0       | 0.015  |
| 0.085             | 1.0           | 0            | 0.5              | 1.0       | 0      |
| 0.1               | 1.0           | 0            | 0.5              | 1.0       | 0      |
| 0.2               | 1.0           | 0            | 0.5              | 1.0       | 0      |

Table 4 - Mahout Random Forest Results, E. coli

Touzain et al. [16] investigated genomic variability at the species level for E. coli, S. aureus and S. pyogenes. It was found that the main type of variability is "microdiversity", the differences in small DNA segments of 20 to 500 bp in size. The S. aureus genomes contained approximately 75% microdiverse loci, with 1.12% insertions and 4.48% deletions per microdiverse loci. The E. coli genomes contained approximately 55% microdiverse loci, with 3.99% insertions and 4.69% deletions per microdiverse loci.

These numbers cannot be directly compared to the value of $p$, but they can be interpreted in the context of results presented. The Random Forest was able to classify successfully actual S. aureus reads, meaning that the variation observed by Touzain et al. is equivalent to a $p$ value of less than the S. aureus critical point of 0.08. This strongly suggests that the Random Forest classifier is capable of classifying actual sequencing projects in addition to these synthetic sequencing projects, but further work is necessary to prove this.
Figure 2 - Top 100 Features, Negative Envelope, $p=0.00$ and $p=0.10$, this time for *E. coli* as the positive class. Note the disruption of separation with increasing noise levels.

5 Conclusions

This work has built on the work completed by Hogan et al. [2] to explore the computational consequences of scaling read classification to a large data set and introducing substantial levels of noise and targeting additional organisms. As before, the approach does not require assembly into contigs or read alignment. Here, the k-mer counts are derived directly from the reads themselves, in contrast to approaches in other domains.

It was found that Random Forests match the classification performance of SVMs when classifying real sequencing projects using a k-mer representation. To allow rapid scaling to large data sets, the read simulator ART was used to generate synthetic sequencing projects from reference genomes. Relief was used to perform feature selection, with a number of custom Python scripts used to automate the data set preparation process.

The computational impacts of scaling to large data sets were investigated in detail. It was found that while the data preparation stage scales linearly, it is by far the most time consuming stage. Data preparation is predominantly parallelisable, but parallel feature selection will be essential for larger data sets. Machine Learning scales linearly with additional sequencing projects and takes an insignificant amount of time when compared to data preparation.

After showing that the software was capable of handling large data sets, while noting these limitations, variation between organisms was investigated. It was found that Mahout's Random Forest was able to successfully classify the synthetic projects up to a well-defined substitution limit above which the classification is rapidly disrupted. These results were observed across a wide range of synthetic sequences and for two distinct target organisms, results that augur well for broader application of the approach. The data set contained multiple *Staphylococcus* strains along with other more distant species, showing that the Random Forest is capable of making both broad and fine-grained classification decisions.
Computationally, the critical contribution lies in the effective use of a simple classifier that may be trained rapidly in parallel and applied through ‘commodity’ parallelism, here Hadoop map-reduce embedded within the Mahout framework. These studies provide a firm basis for further scaling and broader application across the bacterial spectrum, and to eukaryotic sequencing.

References

[1] M. L. Metzker, "Sequencing technologies - the next generation," *Nature Reviews Genetics*, vol. 11, pp. 31-46, 2010.

[2] J. Hogan, P. Holland, A. Holloway, R. Petit and T. Read, "Read Classification for Next Generation Sequencing," in ESANN, Bruges, 2013.

[3] K. Song, J. Ren, G. Reinert, M. Deng, M.S. Waterman and F. Sun, "New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing," *Briefings in Bioinformatics*, 2013.

[4] NCBI, "The Sequence Read Archive," 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/sra.

[5] W. Huang, L. Li, J. Myers and G. Marth, "ART: a next-generation sequencing read simulator," *Bioinformatics*, vol. 28, no. 4, pp. 593-594, 2012.

[6] The Apache Software Foundation, "Apache Mahout," 2013. [Online]. Available: http://mahout.apache.org.

[7] K. Kira and L. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine Learning*, 1992.

[8] E. Chang, K. Zhu, H. Wang, H. Bai, J. Li, Z. Qiu and H. Cui, "PSVM: Parallelizing Support Vector Machines," in *Advances in Neural Information Processing Systems*, 2007.

[9] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," in *OSDI’04: Sixth Symposium on Operating System Design and Implementation*, 2004.

[10] The Apache Software Foundation, "Apache Mahout," 2013. [Online]. Available: http://mahout.apache.org.

[11] The Apache Software Foundation, "Apache Hadoop," 2013. [Online]. Available: http://hadoop.apache.org.

[12] J. Hogan, W. Kelly and F. Newell, "Consensus sigma-70 Promoter Prediction using Hadoop," in *9th IEEE International Conference on e-Science*, Beijing, 2013.

[13] H. Graf, E. Cosatto, L. Bottou, I. Durdanovic and V. Vapnik, "Parallel support vector machines: The cascade SVM," in *Advances in neural information processing systems*, 2004.

[14] M. Friedl and C. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote sensing of environment*, vol. 61, no. 3, pp. 399-409, 1997.

[15] NCBI, "RefSeq: NCBI Reference Sequence Database," 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/RefSeq/.

[16] F. Touzain, E. Denamur, C. Medigue, V. Barbe, M. El Karoui and M. Petit, "Small variable segments constitute a major mtype of diversity of bacterial genomes at the species level," *GenomeBiology*, vol. 11, no. 45, 2010.
Welcome to the 14th Annual International conference on Computational Science, to be held 10th-12th June 2014 in Cairns, Australia. Cairns is on the doorstep of Australian jewels including the Great Barrier Reef and the Daintree rainforest. For more information about Cairns and nearby attractions, see our location description. ICCS 2014 is organized by the University of Queensland, Universiteit van Amsterdam, NTU Singapore and the University of Tennessee.

The International Conference on Computational Science is an annual conference that brings together researchers and scientists from mathematics and computer science as basic computing disciplines, researchers from various application areas who are pioneering computational methods in sciences such as physics, chemistry, life sciences, and engineering, as well as in arts and humanitarian fields, to discuss problems and solutions in the area, to identify new issues, and to shape future directions for research.

Since its inception in 2001, ICCS has attracted increasingly higher quality and numbers of attendees and papers and this year is not an exception. This year we expect over 250 participants. The proceedings series have become a major intellectual resource for computational science researchers and serve to both define and advance the state of the art of the field.

ICCS 2014 in Cairns will be the fourteenth in this series of highly successful conferences. For the previous twelve meetings see: http://iccs2014.ivec.org/previous-iccs.html

The theme for ICCS 2014 is Big Data meets Computational Science to mark the increasing importance of data intensive science. In order to extract meaning from the exponentially increasing amounts of data being gathered, it is imperative to both apply current computational science techniques to data sets, and to develop new processes and algorithms. This conference will be a unique
event focusing on recent developments in: data intensive science for diverse areas of science; scalable scientific algorithms; advanced software tools; computational grids; advanced numerical methods; and novel application areas. ICCS2014 will also feature the important advances in computational science towards exascale computing. ICCS includes work focusing on the application of computational methods in diverse areas including, Computational Biology, Computational Finance, Earth Sciences, Social Sciences and complex systems at large.

ICCS is well known for its excellent line up of keynote speakers and this year is no exception. The keynotes for 2014 are:

- Professor **Vassil Alexandrov**, ICREA Research Professor in Computational Science, Barcelona Supercomputing Centre, Spain
- Professor **Dr Luis Bettencourt**, Santa Fe Institute, New Mexico, USA
- Professor **Peter T. Cummings**, Department of Chemical and Biomolecular Engineering, Vanderbilt University, USA
- **Dan Fey**, Director - Earth, Energy, and Environment Microsoft External Research, Microsoft
- Professor **John Mattick**, Garvan Institute of Medical Research, Sydney, Australia
- Professor **Bob Pressey**, Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Australia
- Professor **Mark Ragan**, Institute for Molecular Bioscience, The University of Queensland, Australia

Besides our excellent keynote speakers, out of the submitted papers to main track and workshops, we selected about 230 high-quality papers for presentation at the conference and publication in the proceedings, published by Elsevier in their Procedia Computer Science series. Submission was very competitive this year and the main track accepted 64 papers from 184 submissions (34% acceptance rate)

ICCS relies strongly on the vital contributions of our workshop organizers to attract high quality papers in many subject areas. We would like to thank all committee members for the main track and the workshops for their contribution to ensure a high standard for the accepted papers. As per every year we would like to thank Elsevier, the conference is organized with their financial and administrative support.

We are proud to note that ICCS is an ERA 2010 A-ranked conference series.

We wish you a successful and enjoyable conference in Cairns.

June 2014

The ICCS 2014 Organizers:
- David Abramson
- Michael Lees
- Valeria V. Krzhizhanovskaya
- Jack Dongarra
- Peter M.A. Sloot
Local organizing committee in Australia

Organizing committee Chair  David Abramson
Organizing committee Members  Jane Carter, Martin Lack, Samantha Hart, Rebecca Moreno

Workshops and Organizers

Multiscale Modelling and Simulation, 11th International Workshop
Valeria Krzhizhanovskaya, Alfons Hoekstra, Derek Groen, Eric Lorenz

5th Workshop on Computational Optimization, Modelling and Simulation
X.S. Yang, S. Koziel, L. Leifsson

Fourth International Workshop on Advances in High-Performance Computational Earth Sciences: Applications and Frameworks
Kengo Nakajima, Huilin Xing

Agent-based simulations, adaptive algorithms and solvers
R. Schaefer, K. Cetnarowicz, V. Calo, D. Pardo, M. Paszynski

Architecture, Languages, Compilation and Hardware support for Emerging ManYcore systems
Loïc Cudennec, Stéphane Louise

Fifth Workshop on Data Mining in Earth System Science
Forrest M. Hoffman, J. Kumar, J. W. Larson, M. D. Mahecha

Dynamic Data Driven Application Systems - DDDAS 2014
C.C. Douglas, A. Patra, A. Cortes

8th Workshop on Computational Chemistry and Its Applications
P. Ramasami

Workshop on Teaching Computational Science
A.B. Shiflet, V. Maxville, Alfredo Tirado-Ramos

Tools for Program Development and Analysis in Computational Science
J. Tao, A. Bode, K. Fürlinger, A. Knüpfer, D. Kranzmüller, J. Volkert, R. Wismüller

Workshop on Cell Based and Individual Based Modelling
J. M. Osborne

Solving Problems with Uncertainties
Vassil Alexandrov

Modeling and Simulation of Large-scale Complex Urban Systems
H. Aydt, M. Berger, X. Li

**Urgent Computing: Computations for Decision Support in Critical Situations**  
A. V. Boukhanovsky, M. Bubak

**Large Scale Computational Physics**  
E. de Doncker, F. Yuasa

**2nd Workshop on Advances in the Kepler Scientific Workflow System and Its Applications**  
I. Altintas, B. Ludaescher

**The Eleventh Workshop on Computational Finance and Business Intelligence**  
Y. Shi, S.Y. Wang, Y.J. Tian

**Bridging the HPC Talent Gap with Computational Science Research Methods**  
E.S. Alexandrova, Vassil Alexandrov

**Mathematical Methods and Algorithms for Extreme Scale**  
Vassil Alexandrov, Jack Dongarra

**Computational Optimisation in the Real World**  
Andrew Lewis, Timoleon Kipouros, Marcus Randall

**7th Workshop on Biomedical and Bioinformatics Challenges for Computer Science**  
M. Cannataro, Pietro Hiram Guzzi, Joakim Sundnes, Rodrigo Weber Dos Santos
Reviewers

D. Abramson
G. Agapito
M. Aldinucci
A. Aleti
V. Alexandrov
E.S. Alexandrova
H. Ali
G. Allen
I. Altintas
S. Ambroszkiewicz
D. Angulo
M. Antolovich
M. Antonieti
J. Antony
H. Aochi
H. Arabnia
P.V. Atherton
M. Auer
H. Aydt
F. Azuaje
E. Bagheri
D.H. Bailey
B. Balis
K. Banas
C. Barrett
R. Bartlett
P. Baruah
D. Bastola
D. Becker
J. Behrens
M. Berger
M. Bernabeu
D. Berrar
M.W. Berry
J. Berthold
J. Betts
S. Bhowmick
S. Blandin
A. Bode
T. Bodisco
B. Boghosian
F. Boniol
B. Bosak
A.V. Boukhanovsky
R. Brito
B.J. Brooks
M. Bruna
M. Bubak
K. Bubendorfer
M. Budka
J. Buisson
K. Burrage
A. Byrski
X. Cai
W. Cai
V. Calo
M. Cannataro
J. Cao
J.C. Carver
K. Cetnarowicz
N. Chandra
A. Chandramowlishwaran
P. Chen
H. Chen
S. Chen
J. Chen
Y. Chen
S.A. Cheong
L.Y. Chew
X. Chi
S.F. Chien
B. Chopard
S. Chuprina
S. Clark
T. Clark
V. Colizza
J. Cooper
A. Cortes
D. Coster
A. Csinkász-Nagy
L. Cudennec
Y. Cui
J. Cunha
L.P. Da Silva Barra
L. Dalcin
S. Date
Y. Davit
M. Dayde
T. Dhaene
G. Di Fatta
S. Diestelhorst
C.H. Ding
G. Dobrowolski
E.H.J. Doncker
J. Dongarra
C.C. Douglas
A. Dragojevic
R. Drezewski
T. Drummond
J. Du
V. Duarte
W. Dubitzky
E. Dugundji
W. Dzwinel
D. Echeverria
N. Emad
C. Engelmann
T. Epperly
C. Erdbrink
J. Fieldsend
L.Jr. Fister
A. Fletcher
J. Flich Cardo
K. Foster
G.C. Fox
C. Froidevaux
K. Fuerlinger
W. Funika
K. Fürlinger
T. Furumura
A.R. Ganguly
L. Garcia-Castillo
A. Garny
F. Gava
Z. Geem
A. Geist
A. Gerbessiotis
O. Ghattas
T. Ghisu
I. Giakiozis
D. Gimenez
J. Glazier
G. Gogniat
B. Gonçalves
B. Goossens
Y. Gorbachev
V. Gramoli
G. Gravvanis
G.A. Gray
C. Greleck
D. Groen
Big Data meets Computational Science... Abramson, Lees, Krzhizhanovskaya, Dongarra, Sloot

L. Gross  C. Guerra  PH Guzzi  U. Hansmann  M. Hardt  W.W. Hargrove  L. Harrison  K. Helmer  T. Hendtlass  V. Hernández  M. Heroux  P. Herrero  D. Hillenbrand  H. Hirst  L. Hluchy  B. Hnatkowska  A. Hoekstra  F.M. Hoffman  D. Howard  R. Hsu  K. Huck  T. Ichimura  A. Inselberg  T. Ishikawa  A. Itkin  S. Ivanov  H. Iwasaki  T. Iwashita  J. Jaros  H. Jin  C. Jin  P. Jöckel  C. Johnson  D. Johnson  X. Ju  H. Kaiser  B.D. Kandhai  E.K. Kansa  A. Karaivanova  C. Kartsaklis  T. Katagiri  W. Kelly  D. Khazanchi  T. Kim  H. Kim  T. Kipouros  A. Knuepfer  A. Knüpfer  M. Koibuchi  V. Korkhov  I. Kotsireas  S. Kovalchuk  S. Koziel  A. Kozionov  D. Kranzlmüller  S. Krishnaswamy  V.V. Krzhizhanovskaya  H. Kugler  J. Kumar  V. Kumar  K. Kurowski  M. Lack  J.W. Larson  N. Le Novre  M. Lees  L. Leifsson  A. Lewis  X. Li  G.T. Lines  C. Liu  M. Lobosco  L. Loew  E. Lorenz  S. Louise  F. Loulegue  P. Lu  B. Ludaescher  E. Luque  S. MacLachlan  M.D. Mahecha  M. Malawski  U. Maran  V. Marangozoa-Martin  S. Margenov  M. Mascagni  L. Maschio  M. Mattavelli  V. Maxville  W. Meira Jr  N. Melab  R. Melnik  J. Michopoulos  R.T. Mills  M. Mirto  H. Mix  K. Mohror  J. Montgomery  I. Moraru  P. Moscato  I. Moser  S. Mostaghim  L. Mountrakis  I. Mozetić  N. Murphy  T. Murphy  P. Murray  O. Mutlu  T. Nakagawa  K. Nakajima  N. Nakasato  P. Navaux  Z. Nemeth  L. Niu  L. Norford  K. Olsen  R. Olsen  S. Orlando  J. M. Osborne  J. Padget  J.P. Papa  M. Paprzycki  D. Pardo  R.S. Parpinelli  A. Paszynska  M. Paszynski  A. Patra  M.S. Pérez  E. Petit  S. Petiton  E. Piriou  J. Pitt Francis  G. Plank  A. Pop  E. Pustulka-Hunt  A. Pyayt  Z. Qi  R. Quax  F.R. Quintana  W. Rachowicz  E. Raffin  P. Raghaven  P. Ramasami  R. Ramirez  O.F. Rana  M. Randall  A. Rau-Chaplin  M. Raulet  A. Rendell  O. Resendis-Antonio  C. Ribbens
| Name                  | Name                        | Name              |
|----------------------|-----------------------------|-------------------|
| M. Riedel            | K. Steinhaeuser             | V. Viswanathan    |
| E. Riviere           | S. Stevenson                | V. Voevodin       |
| Y. Robert            | A. Streit                   | J. Volkert        |
| D. Rodriguez         | H. Sun                      | G. Vozzi          |
| B. Rodriguez         | J. Sundnes                  | J.W. Janneck      |
| T. Ropars            | M. Swain                    | M. Wagner         |
| F. Roux              | C. Swanson                  | D. Walker         |
| K. Rycerz            | M. Swat                     | D. Walker         |
| E. Santos            | R. Tadeusiewicz             | K. Walkowiak      |
| H. Sato              | R. Tagliaferri              | L. Wang           |
| M. Savill            | D. Takahashi                | S.Y. Wang         |
| R. Schaefer          | E. Talbi                    | C. Wang           |
| J. Schaff            | J. Tao                      | B. Wang           |
| B. Schmidt           | O. Tatebe                   | G. Watson         |
| O. Schuetze          | H. Tchelepi                 | R. Weber Dos Santos|
| C. Scoglio           | C. Tedeschi                 | J. Weidendorfer   |
| M. Sensoy            | T. Terlaky                  | R. Wismüller      |
| A. Shafi             | P. Thierry                  | B. Wylie          |
| Y. Shao              | R. Tian                     | R. Wyrzykowski    |
| Y. Shi               | Y.J. Tian                   | H. Xing           |
| A.B. Shiflet         | T.O. Ting                   | X.S. Yang         |
| E.B. Shim            | A. Tirado-Ramos             | C. Yang           |
| T. Shimokawabe       | A. Tiwari                   | F. Yuasa          |
| I. Shin              | P. Trunfio                  | D. Yuen           |
| M. Sicilia           | H. Tufó                     | S. Zasada         |
| F. Silvestri         | P. Turner                   | Q.J. Zhang        |
| J. Sklenar           | S.J. Turner                 | Y. Zhang          |
| P. Sloot             | P. Tvrdik                   | X. Zhao           |
| R. Slota             | H. Usui                     | H. Zheng          |
| S. Smanchat          | D. Van Albada               | Z. Zhou           |
| M. Smolka            | M. Vanderhoef               | X. Zhou           |
| B. Sniezynski        | R.R. Vatsavai               | D. Zmuda          |
| R. Spiteri           | P. Veltri                   | A. Zomaya         |
| P.R. Srivastava      | E.J. Vigmond                | B. Zupan          |
| V. Stankovski        | J. Villà I Freixa           |                  |
# Table of Contents

Big Data meets Computational Science, preface for ICCS 2014 ........................... 1  
*David Abramson, Michael Lees, Valeria Krzhizhanovskaya, Jack Dongarra and Peter M.A. Sloot*

SparseHC: a memory-efficient online hierarchical clustering algorithm ..................... 8  
*Thuy Diem Nguyen, Bertil Schmidt and Chee Keong Kwoh*

Optimizing Shared-Memory Hyperheuristics on top of Parameterized Metaheuristics ...... 20  
*José Matías Cutillas Lozano and Domingo Gimenez*

Tuning Basic Linear Algebra Routines for Hybrid CPU+GPU Platforms ...................... 30  
*Gregorio Bernabé, Javier Cuenca, Domingo Gimenez and Luis-Pedro García*

A portable OpenCL Lattice Boltzmann code for multi- and many-core processor architectures .......................................................... 40  
*Enrico Calore, Sebastiano F. Schifano and Raffaele Tripiccione*

Accelerating Solid-Fluid Interaction using Lattice-Boltzmann and Immersed Boundary Coupled Simulations on Heterogeneous Platforms ........................................ 50  
*Pedro Valero-Lara, Alfredo Pinelli and Manuel Prieto-Matías*

An Empirical Study of Hadoop’s Energy Efficiency on a HPC Cluster ......................... 62  
*Nidhi Tiwari, Santonu Sarkar, Umesh Bellur and Maria Indrawan-Santiago*

Optimal Run Length for Discrete-Event Distributed Cluster-Based Simulations ........... 73  
*Francisco Borges, Albert Gutierrez-Milla, Remo Suppi and Emilio Luque*

A CUDA Based Solution to the Multidimensional Knapsack Problem Using the Ant Colony Optimization ................................................................. 84  
*Henrique Fingler, Edson Cárceles, Henriques Mongelli and Siang Song*

Comparison of High Level FPGA Hardware Design for Solving Tri-Diagonal Linear Systems ........................................................................................................... 95  
*David Warne, Neil Kelson and Ross Hayward*

Blood Flow Arterial Network Simulation with the Implicit Parallelism Library SkelGIS … 102  
*Hélène Coullon, Jose-Maria Fullana, Pierre-Yves Lagrée, Sébastien Limet and Xiaofei Wang*

Triplet Finder: On the Way to Triggerless Online Reconstruction with GPUs for the PANDA Experiment ................................................................. 113  
*Andrew Adinetz, Andreas Herten, Jiří Kraus, Marius Mertens, Dirk Pleiter, Tobias Stockmanns and Peter Wintz*

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2014
A Technique for Parallel Share-Frequent Sensor Pattern Mining from Wireless Sensor Networks
Md Mamunur Rashid, Dr. Iqbal Gondal and Joarder Kamruzzaman

Performance-Aware Energy Saving Mechanism in Interconnection Networks for Parallel Systems
Hai Nguyen, Daniel Franco and Emilo Luque

Handling Data-skew Effects in Join Operations using MapReduce
Mostafa Bamha, Frédéric Loulergue and Mohamad Al Hajj Hassan

Speeding-Up a Video Summarization Approach using GPUs and Multicore-CPU's
Suellen Almeida, Antonio Carlos Nazaré Jr, Arnaldo De Albuquerque Araújo, Guillermo Cámara-Chávez and David Menotti

GPU Optimization of Pseudo Random Number Generators for Random Ordinary Differential Equations
Christoph Riesinger, Tobias Neckel, Florian Rupp, Alfredo Parra Hinojosa and Hans-Joachim Bungartz

Design and Implementation of Hybrid and Native Communication Devices for Java HPC
Bibrak Qamar, Ansar Javed, Mohsan Jameel, Aamir Shafi and Bryan Carpenter

Deploying a Large Petascale System: the Blue Waters Experience
Celso Mendes, Brett Bode, Gregory Bauer, Jeremy Enos, Cristina Beldica and William Kramer

HPC Benchmark Assessment with Statistical Analysis
Fei Xing, Haihang You and Charrn-Da Lu

FPGA-based acceleration of detecting statistical epistasis in GWAS
Lars Wienbrandt, Jan Christian Kässens, Jorge González-Domínguez, Bertil Schmidt, David Ellinghaus and Manfred Schimmler

OS Support for Load Scheduling in Accelerator-based Heterogeneous Systems
Ayman Tarakji, Niels Ole Salscheider and David Hebbeker

Efficient Global Element Indexing for Parallel Adaptive Flow Solvers
Michael Lieb, Tobias Neckel, Hans-Joachim Bungartz and Thomas Schöps

Performance Improvements for a Large-Scale Geological Simulation
David Apostol, Kyle Foerster, Travis Desell and Will Gosnold

Lattice Gas Model for Budding Yeast: A New Approach for Density Effects
Kei-Ichi Tanaka, Takashi Ushimaru, Toshiyuki Hagiwara and Jin Yoshimura

Characteristics of displacement data due to time scale for the combination of Brownian motion with intermittent adsorption
Itsuo Hanasaki, Satoshi Uehara and Satoyuki Kawano

Mechanism of Traffic Jams at Speed Bottlenecks
Wei Liang Quek and Lock Yue Chew

Computing, a powerful tool in flood prediction
Adriana Gaudiani, Emilo Luque, Pablo Garcia, Mariano Re, Marcelo Naiouf and Armando De Giusti
| Title                                                                 | Page |
|----------------------------------------------------------------------|------|
| Evolving Agent-based Models using Complexification Approach          | 310  |
| *Michael Wagner, Wentong Cai, Michael Harold Lees and Heiko Aydt*    |      |
| Discrete modeling and simulation of business processes using event logs | 322  |
| *Ivan Khodyrev and Svetlana Popova*                                   |      |
| Modeling and Simulation Framework for Development of Interactive Virtual Environments | 332  |
| *Konstantin Knyazkov and Sergey Kovalchuk*                           |      |
| Naïve Creature with Fear and Desire Learning to Cross a Highway       | 343  |
| *Anna T. Lawniczak, Jason Ernst and Bruno Di Stefano*                 |      |
| Using interactive 3D game play to make complex medical knowledge more accessible | 354  |
| *Dale Patterson*                                                     |      |
| Measuring the reputation in user-generated-content systems based on health information | 364  |
| *Leila Weitzel, José Palazzo M. De Oliveira and Paulo Quaresma*       |      |
| Spatio-temporal Sequential Pattern Mining for Tourism Sciences       | 379  |
| *Bermingham Luke and Ickjai Lee*                                     |      |
| Benchmarking and Data Envelopment Analysis. An Approach Based on Metaheuristics | 390  |
| *Jose J. Lopez-Espin, Juan Aparicio, Domingo Gimenez and Jesús T. Pastor* |      |
| Improving Collaborative Recommendation via Location-based User-Item Subgroup | 400  |
| *Zhi Qiao, Peng Zhang, Yanan Cao, Chuan Zhou and Li Guo*              |      |
| The influence of network topology on reverse-engineering of gene-regulatory networks | 410  |
| *Alexandru Mizeranschi, Noel Kennedy, Paul Thompson, Huiru Zheng and Werner Dubitzky* |      |
| Maximizing the Cumulative Influence through a Social Network when Repeat Activation Exists | 422  |
| *Chuan Zhou, Peng Zhang, Wenyu Zang and Li Guo*                      |      |
| A Clustering-based Link Prediction Method in Social Networks          | 432  |
| *Fenhua Li, Jing He, Guangyan Huang, Yanchun Zhang and Yong Shi*       |      |
| Discovering Multiple Diffusion Source Nodes in Social Networks        | 443  |
| *Wenyu Zang, Peng Zhang, Chuan Zhou and Li Guo*                      |      |
| Study of the Network Impact on Earthquake Early Warning in the Quake-Catcher Network Project | 453  |
| *Marcos Portnoi, Samuel Schlachter and Michela Taufer*               |      |
| The p-index: Ranking Scientists using Network Dynamics                | 465  |
| *Upul Senanayake, Mahendrarajah Piraveenan and Albert Zomaya*         |      |
| Mining Large-scale Knowledge about Events from Web Text               | 478  |
| *Yanan Cao, Peng Zhang, Jing Guo and Li Guo*                         |      |
| A Technology for BigData Analysis Task Description using Domain-Specific Languages | 488  |
| *Sergey Kovalchuk, Artem Zakharchuk, Jiaqi Liao, Sergey Ivanov and Alexander Boukhanovsky* |      |
A Workflow Application for Parallel Processing of Big Data from an Internet Portal

Pawel Czarnul

A comparative study of scheduling algorithms for the multiple deadline-constrained workflows in heterogeneous computing systems with time windows

Klavdiya Bochenina

Fault-Tolerant Workflow Scheduling Using Spot Instances on Clouds

Deepak Poola, Kotagiri Ramamohanarao and Rajkumar Buyya

On Resource Efficiency of Workflow Schedules

Young Choon Lee, Albert Y. Zomaya and Hyuck Han

Workflow as a Service in the Cloud: Architecture and Scheduling Algorithms

Jianwu Wang, Prakashan Korambath, Ilkay Altintas, Jim Davis and Daniel Crawl

Large Eddy Simulation of Flow in Realistic Human Upper Airways with Obstructive Sleep Apnea

Mingzhen Lu, Yang Liu, Jingying Ye and Haiyan Luo

Experiments on a Parallel Nonlinear Jacobi-Davidson Algorithm

Yoichi Matsuo, Hua Guo and Peter Arbenz

The K computer Operations: Experiences and Statistics

Keiji Yamamoto, Aitsu Uno, Hitoshi Murai, Toshiyuki Tsukamoto, Fumiyoshi Shoji, Shuji Matsui, Ryoichi Sekizawa, Fumichika Sueyasu, Hiroshi Uchiyama, Mitsuo Okamoto, Nobuo Ohgushi, Katsutoshi Takashina, Daisuke Wakabayashi, Yuki Taguchi and Mitsuo Yokokawa

Dendrogram Based Algorithm for Dominated Graph Flooding

Claude Tadonki

HP-DAEMON: High Performance Distributed Adaptive Energy-efficient Matrix-multiplication

Li Tan, Longxiang Chen, Zizhong Chen, Ziliang Zong, Rong Ge and Dong Li

Evaluating the Performance of Multi-tenant Elastic Extension Tables

Haitham Yaish, Madhu Goyal and George Feuerlicht

Finite difference method for solving acoustic wave equation using locally adjustable time-steps

Alexandre Antunes, Regina Leal-Toledo, Otton Filho and Elson Toledo

Identifying Self-Excited Vibrations with Evolutionary Computing

Christiaan Erbrink and Valeria Krzhizhanovskaya

Rendering of Feature-Rich Dynamically Changing Volumetric Datasets on GPU

Martin Schreiber, Atanas Atanasov, Philipp Neumann and Hans-Joachim Bungartz

Motor learning in physical interfaces for computational problem solving

Rohan McAdam

Change Detection and Visualization of Functional Brain Networks using EEG Data

R Vijayalakshmi, Naga Dasari, Nanda Nandagopal, R Subhiksha, Bernadine Cocks, Nabaraj Dahal and M Thilaga
Visual Analytics of Topological Higher Order Information for Emergency Management based on Tourism Trajectory Datasets
Ye Wang, Kyungmi Lee and Ickjai Lee

Modulight: A Framework for Efficient Dynamic Interactive Scientific Visualization
Sébastien Limet, Millian Poquet and Sophie Robert

Visualization of long-duration acoustic recordings of the environment
Michael Towsey, Liang Zhang, Mark Cottman-Fields, Jason Wimmer, Jinglan Zhang and Paul Roe

A computational science agenda for programming language research
Dominic Orchard and Andrew Rice

Restrictions in model reduction for polymer chain models in dissipative particle dynamics
Nicolas Moreno, Suzana Nunes and Victor M. Calo

Simulation platform for multiscale and multiphysics modeling of OLEDs
Maria Bogdanova, Sergey Belousov, Ilya Valuev, Andrey Zakirov, Mikhail Okun, Denis Shirkabagkin, Vasily Cherkov, Petr Tokar, Andrey Knizhnik, Boris Potapkin, Alexander Bagaturyants, Ksenia Komarova, Mikhail Strikhanov, Alexey Tishchenko, Vladimir Nikitenko, Vasily Sukharev, Natalia Sannikova and Igor Morozov

Computational Optimization, Modelling and Simulation: Past, Present and Future
Xin-She Yang, Slawomir Koziel and Leifur Leifsson

Minimizing Inventory Costs for Capacity-Constrained Production using a Hybrid Simulation Model
John Betts

Nested Space Mapping Technology for Expedite EM-driven Design of Compact RF/microwave Components
Adrian Bekasiewicz, Slawomir Koziel, Piotr Kurgan and Leifur Leifsson

Solution of the wave-type PDE by numerical damping control multistep methods
Elisabete Alberdi Celaya and Juan José Anza Aguirrezabala

Low-Cost EM-Simulation-Driven Multi-Objective Optimization of Antennas
Adrian Bekasiewicz, Slawomir Koziel and Leifur Leifsson

Computation on GPU of Eigenvalues and Eigenvectors of a Large Number of Small Hermitian Matrices
Alain Cosnuau

Fast Low-fidelity Wing Aerodynamics Model for Surrogate-Based Shape Optimization
Leifur Leifsson, Slawomir Koziel and Adrian Bekasiewicz

COFADMM: A Computational features selection with Alternating Direction Method of Multipliers
Mohammed Elanbari, Sidra Alam and Halima Bensmail

Preference-Based Fair Resource Sharing and Scheduling Optimization in Grid VOs
Victor Toporkov, Anna Toporkova, Alexey Tselishchev, Dmitry Yemelyanov and Petr Potekhin
Variable Neighborhood Search Based Set covering ILP model for the Vehicle Routing Problem with time windows ......................................................... 844
    Amine Dhahri, Kamel Zidi and Khaled Ghedira

A physics-based Monte Carlo earthquake disaster simulation accounting for uncertainty in building structure parameters ........................................ 855
    Shunsuke Homma, Kohei Fujita, Tsuyoshi Ichimura, Muneo Hori, Seckin Citak and Takane Hori

A quick earthquake disaster estimation system with fast urban earthquake simulation and interactive visualization ........................................... 866
    Kohei Fujita, Tsuyoshi Ichimura, Muneo Hori, M. L. L. Wijerathne and Seizo Tanaka

Several hundred finite element analyses of an inversion of earthquake fault slip distribution using a high-fidelity model of the crustal structure .......... 877
    Ryoichiro Agata, Tsuyoshi Ichimura, Kazuro Hirahara, Mamoru Hyodo, Takane Hori and Muneo Hori

An out-of-core GPU approach for Accelerating Geostatistical Interpolation ............ 888
    Victor Allombert, David Michea, Fabrice Dupros, Christian Bellier, Bernard Bourgine, Hideo Aochi and Sylvain Jubertie

Mesh generation for 3D geological reservoirs with arbitrary stratigraphic surface constraints ........................................................... 897
    Huilin Xing and Yan Liu

Application-specific I/O Optimizations on Petascale Supercomputers .................. 910
    Efecan Poyraz, Heming Xu and Yifeng Cui

Performance evaluation and case study of a coupling software ppOpen-MATH/MP ...... 924
    Takashi Arakawa, Takahiro Inoue and Masaki Sato

Implementation and Evaluation of an AMR Framework for FDM Applications ........ 936
    Masaharu Matsumoto, Futoshi Mori, Satoshi Ohshima, Hideyuki Jitsumoto, Takahiro Katagiri and Kengo Nakajima

Dynamic Programming Algorithm for Generation of Optimal Elimination Trees for Multi-Frontal Direct Solver over h Refined Grids ...................... 947
    Hassan Aboueisha, Mikhail Moshkov, Victor Calo, Maciej Paszynski, Damian Goik and Konrad Jopek

Graph grammar based multi-thread multi-frontal direct solver with Galois scheduler ... 960
    Damian Goik, Konrad Jopek, Maciej Paszynski, Andrew Lenharth, Donald Nguyen and Keshav Pingali

Automatically Adapted Perfectly Matched Layers for Problems with High Contrast Materials Properties ......................................................... 970
    Julen Alvarez-Aramberri, David Pardo, Helene Barucq and Elisabete Alberdi Celaya

Modeling phase-transitions using a high-performance, Isogeometric Analysis framework ... 980
    Philippe Vignal, Lisandro Dalcin, Nathan Collier and Victor Calo

Micropolar Fluids using B-spline DivergenceConforming Spaces .......................... 991
    Adel Sarmiento, Daniel Garcia, Lisandro Dalcin, Nathan Collier and Victor Calo
Hypergraph grammar based adaptive linear computational cost projection solvers for
two and three dimensional modeling of brain ............................................. 1002
  Damian Goik, Marcin Sieniek, Maciej Woźniak, Anna Paszyńska and Maciej Paszynski
Implementation of an adaptive BDF2 formula and comparison with the MATLAB ode15s 1014
  Elisabete Alberdi Celaya, Juan José Anza Aguirrazabala and Panagiotis Chatzipantelidis
Fast graph transformation based direct solver algorithm for regular three dimensional
grids ............................................................................................................ 1027
  Marcin Sieniek
Agent-based Evolutionary Computing for Difficult Discrete Problems .................. 1039
  Michal Kowol, Aleksander Byrski and Marek Kisiel-Dorohinicki
Translation of graph-based knowledge representation in multi-agent system ............... 1048
  Leszek Kotulski, Adam Sedziwy and Barbara Strug
Agent-based Adaptation System for Service-Oriented Architectures Using Supervised
Learning ........................................................................................................ 1057
  Bartłomiej Snieżynski
Generation-free Agent-based Evolutionary Computing ........................................ 1068
  Daniel Krzywicki, Jan Stypka, Piotr Anielski, Łukasz Faber, Wojciech Turek,
  Aleksander Byrski and Marek Kisiel-Dorohinicki
Hypergraph grammar based linear computational cost solver for three dimensional grids
with point singularities ............................................................................... 1078
  Piotr Gurgul, Anna Paszynska and Maciej Paszynski
A Linear Complexity Direct Solver for H-adaptive Grids With Point Singularities ...... 1090
  Piotr Gurgul
C: C with Process Network Extensions for Embedded Manycores ....................... 1100
  Thierry Goubier, Damien Courousse and Selma Azaiez
Application-Level Performance Optimization: A Computer Vision Case Study on
STHORM ................................................................................................. 1113
  Vítor Schwambach, Sébastien Cleyet-Merle, Alain Issard and Stéphane Mancini
Generating Code and Memory Buffers to Reorganize Data on Many-core Architectures
  Loïc Cudennec, Paul Dubrulle, François Galea, Thierry Goubier and Renaud Sirdey
Self-Timed Periodic Scheduling For Cyclo-Static DataFlow Model ........................ 1134
  Dkhil Ep. Jemal Amira, Xuankhanh Do, Stephane Louise, Dubrulle Paul and Christine
  Rochang
Stochastic Parameterization to Represent Variability and Extremes in Climate Modeling
  Roisin Langan, Richard Archibald, Matthew Plumlee, Salid Mahajan, Daniel Ricciuto,
  Cheng-En Yang, Rui Mei, Jiafu Mao, Xiaoying Shi and Joshua Fu
Integration of artificial neural networks into operational ocean wave prediction models
for fast and accurate emulation of exact nonlinear interactions .......................... 1156
  Ruslan Puscasu
Control of Artificial Swarms with DDDAS ............................................... 1171
  Robert McCune and Greg Madey

vi
Multifidelity DDDAS Methods with Application to a Self-Aware Aerospace Vehicle

Doug Allaire, David Kordonowy, Marc Leceerf, Laura Mainini and Karen Willcox

Model Based Design Environment for Data-Driven Embedded Signal Processing Systems

Kishan Sudusinghe, Inkeun Cho, Mihaela van der Schaar and Shweta Bhattacharyya

A Dynamic Data Driven Application System for Vehicle Tracking

Richard Fujimoto, Angshuman Guin, Michael Hunter, Haesun Park, Ramakrishnan Kannan, Gaurav Kanitkar, Michael Milholen, Sabra Neal and Philip Pecher

Towards a Dynamic Data Driven Wildfire Behavior Prediction System at European Level

Tomás Artés, Andrés Cencerrado, Ana Cortes, Tomas Margalef, Darío Rodríguez, Thomas Petrolaquis and Jesus San Miguel

Fast Construction of Surrogates for UQ Central to DDDAS – Application to Volcanic Ash Transport

A. K. Patra, E. R. Stefanescu, R. M. Madankan, M. I Bursik, E. B. Pitman, P. Singla, T. Singh and P. Webley

A Dynamic Data-driven Decision Support for Aquaculture Farm Closure

Md. Sumon Shahriar and John McCulloch

An Open Framework for Dynamic Big-Data-Driven Application Systems (DBDDAS) Development

Craig C. Douglas

A posteriori error estimates for DDDAS inference problems

Vishwas Hebbur Venkata Subba Rao and Adrian Sandu

Mixture Ensembles for Data Assimilation in Dynamic Data-Driven Environmental Systems

Piyush Tagade, Hansjorg Seybold and Sai Ravela

Optimizing Dynamic Resource Allocation

Lucas Krakow, Louis Rabiet, Yun Zou, Guillaume Iooss, Edwin Chong and Sanjay Rajopadhye

A Dataflow Programming Language and Its Compiler for Streaming Systems

Haitao Wei, Stephane Zuckerman, Xiaoming Li and Guang Gao

Static versus Dynamic Data Information Fusion analysis using DDDAS for Cyber Security Trust

Erik Blasch, Youssif Al-Nashif and Salim Hariri

Dynamic Data Driven Crowd Sensing Task Assignment

Layla Pournajaf, Li Xiong and Vaidy Sunderam

Context-aware Dynamic Data-driven Pattern Classification*

Shashi Phoha, Nurali Virani, Prithi Chattopadhyay, Soumalya Sarkar, Brian Smith and Asok Ray

Correlation between Franck-Condon Factors and Average Internuclear Separations for Diatomics Using the Fourier Grid Hamiltonian Method

Mayank Kumar Dixit, Abhishek Jain and Bhalachandra Laxmanrao Tembe
Using hyperheuristics to improve the determination of the kinetic constants of a chemical reaction in heterogeneous phase

José Matías Cutillas Lozano and Domingo Gimenez

A Computational Study of 2-Selenobarbituric Acid: Conformational Analysis, Enthalpy of Formation, Acidity and Basicity

Rafael Notario

Origin of the Extra Stability of Alloxan. A Computation Study

Saadullah Aziz, Rifaat Hilal, Basmah Allehyani and Shabaan Elrouby

The Impact of p-orbital on Optimization of ReH7(PMe3)2 Compound

Nnenna Elechi, Daniel Tran, Mykala Taylor, Odaro Adu and Huajun Fan

Exploring the Conical Intersection Seam in Cytosine: A DFT and CASSCF Study

Rifaat Hilal, Saadullah Aziz, Shabaan Elrouby and Walid Hassan

An Introduction to Agent-Based Modeling for Undergraduates

Angela Shiflet and George Shiflet

Computational Science for Undergraduate Biologists via QUT.Bio.Excel

Lawrence Buckingham and James Hogan

A multiple intelligences theory-based 3D virtual lab environment for digital systems teaching

Toni Amorim, Norian Marranghello, Alexandre C.R. Silva, Aledir S. Pereira and Lenadro Tapparo

Exploring Rounding Errors in Matlab using Extended Precision

Dina Tsarapkina and David Jeffrey

Double-Degree Master’s Program in Computational Science: Experiences of ITMO University and University of Amsterdam

Alexey Dukhanov, Valeria Krzhizhanovskaya, Anna Bilyatdinova, Alexander Boukhanovsky and Peter Sloot

A High Performance Computing Course Guided by the LU Factorization

Gregorio Bernabé, Javier Cuenca, Luis P. García, Domingo Gimenez and Sergio Rivas-Gomez

Teaching High Performance Computing using BeesyCluster and Relevant Usage Statistics

Pawel Czarnul

High Performance Message-Passing InfiniBand Communication Device for Java HPC

Omar Khan, Mohsan Jameel and Aamir Shafi

A High Level Programming Environment for Accelerator-based Systems

Luiz Derose, Heidi Poxon, James Beyer and Alistair Hart

Supporting relative debugging for large-scale UPC programs

Minh Ngoc Dinh, David Abramson, Jin Chao, Bob Moench, Andrew Gontarek and Luiz Derose

Near Real-time Data Analysis of Core-Collapse Supernova Simulations With Bellerophon

E. J. Lingerfelt, O. E. B. Messer, S. S. Desai, C. A. Holt and E. J. Lentz
Toward Better Understanding of the Community Land Model within the Earth System Modeling Framework ................................................................. 1515  
Dali Wang, Joseph Schuchart, Tomislav Janjusic, Frank Winkler, Yang Xu and Christos Kartsaklis

Detecting and visualising process relationships in Erlang  ........................................ 1525  
Melinda Tóth and István Bozó

Wind field uncertainty in forest fire propagation prediction ........................................ 1535  
Gemma Sanjuan, Carlos Brun, Tomas Margalef and Ana Cortes

A Framework for Evaluating Skyline Query over Uncertain Autonomous Databases ...... 1546  
Nurul Husna Mohd Saad, Hamidah Ibrahim, Ali Amer Alwan, Fatimah Sidi and Razali Yaakob

Efficient Data Structures for Risk Modelling in Portfolios of Catastrophic Risk Using MapReduce ................................................................. 1557  
Andrew Rau-Chaplin, Zhimin Yao and Norbert Zeh

Argumentation Approach and Learning Methods in Intelligent Decision Support Systems in the Presence of Inconsistent Data ................................. 1569  
Vadim N. Vagin, Marina Fomina and Oleg Morosin

Enhancing Monte Carlo Preconditioning Methods for Matrix Computations ............. 1580  
Janko Strassburg and Vassil Alexandrov

Analysing the Effectiveness of Wearable Wireless Sensors in Controlling Crowd Disasters  ......................................................... 1590  
Teo Yu Hui Angela, Vaisagh Viswanathan, Michael Lees and Wentong Cai

Individual-Oriented Model Crowd Evacuations Distributed Simulation ........................ 1600  
Albert Gutierrez-Milla, Francisco Borges, Remo Suppi and Emilio Luque

Simulating Congestion Dynamics of Train Rapid Transit using Smart Card Data ........ 1610  
Nasri Othman, Erika Fille Legara, Vicknesh Selvam and Christopher Monterola

A method to ascertain rapid transit systems’ throughput distribution using network analysis ...................................................................................... 1621  
Muhamad Azfar Ramli, Christopher Monterola, Gary Kee Khoon Lee and Terence Gih Guang Hung

Fast and Accurate Optimization of a GPU-accelerated CA Urban Model through Cooperative Coevolutionary Particle Swarms ........................................ 1631  
Ivan Blecic, Arnaldo Cecchini and Giuseppe A. Trunfio

High Performance Computations for Decision Support in Critical Situations: Introduction to the Third Workshop on Urgent Computing ................................. 1644  
Alexander Boukhanovsky and Marian Bubak

Personal decision support mobile service for extreme situations ................................ 1646  
Vladislav A. Karbovskii, Daniil V. Voloshin, Kseniia A. Puzyrev and Aleksandr S. Zagarskikh

Evaluation of in-vehicle decision support system for emergency evacuation ............ 1656  
Sergei Ivanov and Konstantin Knyazkov
Problem solving environment for development and maintenance of St. Petersburg's Flood Warning System ................................................................. 1667
  Sergey Kosukhin, Anna Kalyuzhnaya and Denis Nasonov

Hybrid scheduling algorithm in early warning ........................................... 1677
  Denis Nasonov and Nikolay Butakov

On-board Decision Support System for Ship Flooding Emergency Response .......... 1688
  Jose Varela, Jose Rodrigues and Carlos Guedes Soares

Development of lattice QCD simulation code set “Bridge++” on accelerators ........ 1701
  Shinji Motoki, Shinya Aoki, Tatsumi Aoyama, Kazuyuki Kanaya, Hideo Matsufuru,
  Yusuke Namekawa, Hidekatsu Nemura, Yusuke Taniguchi, Satoru Ueda and Naoya Ukita

GPGPU Application to the Computation of Hamiltonian Matrix Elements between Non-orthogonal Slater Determinants in the Monte Carlo Shell Model ............... 1711
  Tomoaki Togashi, Noritaka Shimizu, Yutaka Utsuno, Takashi Abe and Takaharu Otsuka

Design and Implementation of Kepler Workflows for BioEarth ...................... 1722
  Tristan Mullis, Mingliang Liu, Ananth Kalyanaraman, Joseph Vaughan, Christina Tague and Jennifer Adam

Tools, methods and services enhancing the usage of the Kepler-based scientific workflow framework ................................................................. 1733
  Marcin Plociennik, Szymon Winczewski, Paweł Ciecielag, Frederic Imbeaux, Bernard Guillermo, Philippe Huynh, Michał Owsiak, Piotr Spyra, Thierry Aniel, Bartek Palak, Tomasz Żok, Wojciech Pych and Jarosław Rybicki

Progress towards automated Kepler scientific workflows for computer-aided drug discovery and molecular simulations .............................................. 1745
  Pek U. Ieong, Jesper Sørensen, Prasanth L. Vemu, Celia W. Wong, Özlem Demir,
  Nadya P. Williams, Jianwu Wang, Daniel Crawl, Robert V. Swift, Robert D. Malmström, Ilkay Altintas and Rommie E. Amaro

Flexible approach to astronomical data reduction workflows in Kepler ............ 1756
  Paweł Ciecielag, Marcin Plociennik, Piotr Spyra, Michal Urbaniak, Tomasz Żok and Wojciech Pych

Identifying Information Requirement for Scheduling Kepler Workflow in the Cloud ...... 1762
  Sucha Smanchat and Kanchana Viriyapant

Twin Support Vector Machine in Linear Programs ...................................... 1770
  Dewei Li and Yingjie Tian

Determining the time window threshold to identify user sessions of stakeholders of a commercial bank portal ......................................................... 1779
  Jozef Kapusta, Michal Munk, Peter Suec and Anna Pilkova

Historical Claims Data Based Hybrid Predictive Models for Hospitalization ........ 1791
  Chengcheng Liu and Yong Shi

Research on the construction of macro assets price index based on support vector machine 1801
  Ping Liu, Jianmin Sun, Liying Han and Bo Wang
In Need of Partnerships  An Essay about the Collaboration between Computational
Sciences and IT Services ................................................................. 1816
   Anton Frank, Ferdinand Jamitzky, Helmut Satzger and Dieter Kranzlmüller

Development of Multiplatform Adaptive Rendering Tools to Visualize Scientific
Experiments ............................................................................. 1825
   Konstantin Ryabinin and Svetlana Chuprina

Education 2.0: Student Generated Learning Materials through Collaborative Work ...... 1835
   Raul Ramirez-Velarde, Raul Perez-Cazares, Nia Alexandrov and Jose Jesus
   Garcia-Rueda

Measuring Business Value of Learning Technology Implementation in Higher Education
Setting ................................................................................... 1846
   Nia Alexandrov

Fast Iterative Method in solving Eikonal equations : a multi-level parallel approach...... 1859
   Florian Dang and Nahid Emad

A one-step Steffensen-type method with super-cubic convergence for solving nonlinear
equations ................................................................................ 1870
   Zhongli Liu

A Parallel Implementation of Singular Value Decomposition for Video-on-Demand
Services Design Using Principal Component Analysis ............................. 1876
   Raul Ramirez-Velarde, Martin Roderus, Carlos Barba-Jimenez and Raul Perez-Cazares

Scalable Stochastic and Hybrid Methods and Algorithms for Extreme Scale Computing...1888
   Vassil Alexandrov

Extending the Front: Designing RFID Antennas using Multiobjective Differential
Evolution with Biased Population Selection ..................................... 1893
   James Montgomery, Marcus Randall and Andrew Lewis

Local Search Enabled Extremal Optimisation for Continuous Inseparable
Multi-objective Benchmark and Real-World Problems ........................... 1904
   Marcus Randall, Andrew Lewis, Jan Hettenhausen and Timoleon Kipouros

A Web-Based System for Visualisation-Driven Interactive Multi-Objective Optimisation..1915
   Jan Hettenhausen, Andrew Lewis and Timoleon Kipouros

A Hybrid Harmony Search Algorithm for Solving Dynamic Optimisation Problems ...... 1926
   Ayad Turky, Salwani Abdullah and Nasser Sabar

Constraint Programming and Ant Colony System for the Component Deployment
Problem .......................................................... 1937
   Dhananjay Thiruvady, I. Moser, Aldeida Aleti and Asef Nazari

Electrical Power Grid Network Optimisation by Evolutionary Computing ............... 1948
   John Oliver, Timoleon Kipouros and Mark Savill

Integrating Genetic and Clinical Data: proposal of an architecture and data model ...... 1959
   Giuseppe Tradigo, Claudia Veneziano, Pierangelo Veltri and Sergio Greco
Mining Association Rules from Gene Ontology and Protein Networks: Promises and Challenges ................................. Pietro Hiram Guzzi, Marianna Milano and Mario Cannataro

Automated Microalgae Image Classification .................................. Sansoen Promdaen, Pakaket Wattuya and Nuttha Sanevas

A Clustering Based Method Accelerating Gene Regulatory Network Reconstruction ...... Georgios Dimitrakopoulos, Ioannis Maraziotis, Kyriakos Sgarbas and Anastasios Bezerianos

Large Scale Read Classification for Next Generation Sequencing ...................... James Hogan and Timothy Peut

Computation of Filtering Functions for Cryptographic Applications............... Amparo Fuster-Sabater

Cachaça Classification Using Chemical Features and Computer Vision ............... Bruno Urbano Rodrigues, Ronaldo Martins Da Costa, Rogério Lopes Salvini and Anderson Da Silva Soares

Web- and Cloud-based Software Infrastructure for Materials Design ................. Janos Sallai, Gergely Varga, Sara Toth, Christopher Iacovella, Christoph Klein, Clare McCabe, Akos Ledeczi and Peter Cummings

Automated estimation and analysis of pulmonary function test parameters from spirometric data for respiratory disease diagnostics ...................... Ritaban Dutta

Distance-Based High-Frequency Trading ........................................ Travis Felker, Vadim Mazalov and Stephen Watt

Multi-Scale Foreign Exchange Rates Ensemble for Classification of Trends in Forex Market Hossein Talebi, Winsor Hoang and Marina L. Gavrilova

Evaluating Parallel Programming Tools to Support Code Development for Accelerators… Rebecca Hartman-Baker, Valerie Maxville and Daniel Grimwood

A Multi-layer Event Detection Algorithm for Detecting Global and Local Hot Events in Social Networks .................................................. Zhicong Tan

Low-dimensional visualization of experts preferences in urgent group decision making under uncertainty .............................. Iván Palomares and Luis Martínez

A Semi-discretized Numerical Method for Solving One Type of Singular Integro-differential Equation Containing Derivatives of the Possible Delay States Shihchung Chiang

A Hybrid MPI+OpenMP Solution of the Distributed Cluster-Based Fish Schooling Simulator ................................................................. Francisco Borges, Albert Gutierrez-Milla, Remo Suppi and Emilio Luque

Hierarchical emulation and data assimilation into the sediment transport model ........ Nugzar Maryvelashvili, Eddy Campbell, Laurence Murray and Emlyn Jones
Cyclic hybrid ow-shop scheduling problem with machine setups
Wojciech Bozejko, Łukasz Gniewkowski, Jaroslaw Pempera and Mieczysław Wodecki

Enabling Global Experiments with Interactive Reconfiguration and Steering by Multiple Users
Luis Assuncao and Jose Cunha

Numerical Optimization Technique for Optimal Design of the n Grooves Surface Plasmon Grating Coupler
Carmen Caiseda, Igor Griva, Luis Martinez, Kyle Shaw and Dan Weingarten

The Design and Implementation of a GPU-enabled Multi-Objective Tabu-Search intended for Real World and High-Dimensional Applications
Christos Tsotskas, Timoleon Kipouros and Mark Savill

Using Kepler for Tool Integration in Microarray Analysis Workflows
Zhuohui Gan, Jennifer Stowe, Andrew McCulloch, Alex Zambon and Ilkay Altintas

Multi-tenant Elastic Extension Tables Data Management
Haitham Yaish, Madhu Goyal and George Feuerlicht

The container problem in a torus-connected cycles network
Antoine Bossard and Keiichi Kaneko

The Knapsack Problem with Three Practical Constraints
Rainne Florisbelo Gonçalves and Thiago Alves De Queiroz

Autonomous Framework for Sensor Network Quality Annotation: Maximum Probability Clustering Approach
Ritaban Dutta

A Performance Model for OpenMP Memory Bound Applications in Multisocket Systems
César Allande, Josep Jorba, Anna Sikora and Eduardo Cesar

Image Noise Removal on Heterogeneous CPU-GPU Configurations
Josep Arnal, M. Guadalupe Sánchez, Vicente Vidal and Anna Vidal

A Faster Parallel Algorithm for Matrix Multiplication on a Mesh Array
Tong-Wook Shinn, Sung Eun Bae and Tadao Takaoka

Cluster-based communication and load balancing for simulations on dynamically adaptive grids
Martin Schreiber and Hans-Joachim Bungartz

Deploying Kepler Workflows as Services on a Cloud Infrastructure for Smart Manufacturing
Prakashan Korambath, Jianwu Wang, Ankur Kumar, Lorin Hochstein, Brian Schott, Robert Graybill, Michael Baldea and Jim Davis

A Fine-grained Approach for Power Consumption Analysis and Prediction
Claude Tadonki

Performance of Unidirectional Hierarchization for Component Grids Virtually Maximized
Philipp Hupp
The WorkWays problem solving environment
Hoang Nguyen, David Abramson and Timoleon Kipouros

EPIK-a Workflow for Electron Tomography in Kepler
Ruijuan Chen, Xiaohua Wan, Ilkay Altintas, Jianwu Wang, Daniel Crawl, Sébastien Phan, Albert Lawrence and Mark Ellisman

Productivity frameworks in big data image processing computations - creating photographic mosaics with Hadoop and Scalding
Piotr Soul and Tomasz Bednarz

Novel Concepts for Realizing Neural Networks as Services in the Sky
Altaf Ahmad Huqqani, Erich Schikuta and Erwin Mann

Complex Network Modeling for Maritime Search and Rescue Operations
Alexey Bezgodov and Dmitrii Esin

Data Centric Framework for Large-scale High-performance Parallel Computation
Kenji Ono, Yasuhiro Kawashima and Tomohiro Kawanabe

Development of a Computational Framework for Block-Based AMR Simulations
Hideyuki Usui, Akihide Nagara, Masanori Nunami and Masaharu Matsumoto

A Resource Efficient Big Data Analysis Method for the Social Sciences: the case of global IP activity
Klaus Ackermann and Simon D. Angus

Impact of I/O and Data Management in Ensemble Large Scale Climate Forecasting Using EC-Earth3
Muhammad Asif, Andrés Cencerrado, Oriol Mula-Valls, Domingo Manubens, Francisco Doblas-Reyes and Ana Cortés

Hybrid Message Logging. Combining advantages of Sender-based and Receiver-based approaches
Hugo Daniel Meyer, Dolores Rexachs and Emilo Luque

Pseudorandom Number Generation in the Context of a 3D Simulation Model for Tissue Growth
Belgacem Ben Youssef and Rachid Sammouda

Evolutionary simulation of complex networks structures with specific topological properties
Victor Kashirin

Modeling and Visualization individual and collective opinions towards extremism in a society
Vinicius Nonnemacher, Luiz Paulo Luna de Oliveira, Marta Becker Villamil and Bardo E. J. Bodmann

POSH: Paris OpenSHMEM A High-Performance OpenSHMEM Implementation for Shared Memory Systems
Camille Coti

Online Collaborative Environment for Designing Complex Computational Systems
Miklos Maroti, Robert Kereskenyi, Tamas Kecskes, Peter Volgyesi and Akos Ledeczi
Computation of ECG signal features using MCMC modelling in software and FPGA reconfigurable hardware ................................................................. 2442
  Timothy Bodisco, Jason D’Netto, Neil Kelson, Jasmine Banks and Ross Hayward

Node assortativity in complex networks: An alternative approach ...................... 2449
  Upul Senanayake, Mahendra Piraveenan, Dharshana Kasthuriratna and Gnana Thedchanamoorthy

Social networks mining for analysis and modeling drugs usage .............................. 2462
  Andrei Yakushev and Sergey Mityagin

Design Virtual Learning Labs for Courses in Computational Science with use of Cloud Computing Technologies ................................................................. 2472
  Alexey Dukhanov, Maria Karpova and Klavdiya Bochenina

The Framework for Problem Solving Environments in Urban Science ..................... 2483
  Aleksandr Zagarskikh, Andrey Karsakov and Timofey Tchurov

Interpolation of Sensory Data in the Presence of Obstacles .................................... 2496
  Dongzhi Zhang and Ickjai Lee

Domain Ontologies Integration for Virtual Modelling and Simulation Environments ...... 2507
  Pavel Smirnov, Sergey Kovalchuk and Alexey Dukhanov

Characteristics of Dynamical Phase Transitions for Noise Intensities ....................... 2515
  Muyoung Heo, Jong-Kil Park and Kyungsik Kim
"Big Data meets Computational Science"

The International Conference on Computational Science is an annual conference that brings together researchers and scientists from mathematics and computer science as basic computing disciplines, researchers from various application areas who are pioneering computational methods in sciences such as physics, chemistry, life sciences, and engineering, as well as in arts and humanitarian fields, to discuss problems and solutions in the area, to identify new issues, and to shape future directions for research.

ICCS 2014 in Cairns, Queensland, will be the fourteenth in this series of highly successful conferences. Cairns is on the doorstep of Australian jewels including the Great Barrier Reef and the Daintree rainforest. For more information about Cairns and nearby attractions, see our location description.

Since its inception in 2001, ICCS has attracted increasingly higher quality and numbers of attendees and papers. Average attendance each year is about 350 participants. The proceedings series have become a major intellectual resource for computational science researchers and serve to both define and advance the state of the art of the field. An archive of the previous meetings is available through the Previous ICCS conferences page.

Out of the submitted papers to main track and workshops, we will select some 30% high-quality papers for presentation at the conference and publication in the proceedings. These are published by Elsevier in open-access Procedia Computer Science series and indexed by Scopus, ScienceDirect, Thomson Reuters Conference Proceedings Citation (former ISI Proceedings) - an integrated index within Web of Science. The papers will contain linked references, XML versions and citable DOI numbers.

ICCS is well known for its excellent line up of keynote speakers. The keynotes for 2014 are:

- Professor Vassil Alexandrov, ICREA Research Professor in Computational Science, Barcelona Supercomputing Centre, Spain
- Professor Dr Luis Bettencourt, Santa Fe Institute, New Mexico, USA
- Professor Professor Peter T. Cummings, Department of Chemical and Biomolecular Engineering, Vanderbilt University, USA
- Dan Fay, Director - Earth, Energy, and Environment Microsoft External Research, Microsoft
- Dr Warren Kaplan, Garvan Institute of Medical Research, Sydney, Australia
- Professor Bob Pressey, Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Australia
- Professor Mark Ragan, Institute for Molecular Bioscience, The University of Queensland, Australia

The theme for ICCS 2014 is Big Data meets Computational Science to mark the increasing importance of data intensive science. In order to extract meaning from the exponentially increasing amounts of data being gathered, it is imperative to both apply current computational science techniques to data sets, and to develop new processes and algorithms. This conference will be a unique event focusing on recent developments in: data intensive science for diverse areas of science; scalable scientific algorithms; advanced software tools; computational grids; advanced numerical methods; and novel application areas. These innovative novel models, algorithms and tools drive new science through efficient application in areas such as physical systems, computational and systems biology, environmental systems, finance, and others.

We look forward to welcoming you to this exciting event in Australia!

ICCS 2014 is organised by

The University of Queensland
Research Computing Centre

Elsevier

University of Amsterdam

The University of Tennessee, Knoxville

http://www.iccs-meeting.org/iccs2014/
ICCS 2014 invites original contributions on all topics related to Computational Science, including, but not limited to:

- Scientific Computing
- Problem Solving Environments
- Advanced Numerical Algorithms
- Complex Systems: Modeling and Simulation
- Hybrid Computational Methods
- Web- and Grid-based Simulations and Computing
- Parallel and Distributed Computing
- Advanced Computing Architectures and New Programming Models
- Visualization and Virtual Reality as Applied to Computational Science
- Applications of Computation as a Scientific Paradigm
- New Algorithmic Approaches to Computational Kernels and Applications
- Computational Humanities
- Education in Computational Science
- Large Scale Scientific Instruments
- Computational Sociology

As the conference theme this year is "Big Data meets Computational Science", workshops that highlight responses to the challenges of big data in various scientific disciplines will be particularly welcomed.

The submitted paper must be camera-ready and formatted according to the rules of Procedia Computer Science. Please use this file for a Latex template plus instructions and click here for an MS word template file. Submission implies the willingness of at least one of the authors to register and present the paper. PostScript and source versions of your paper must be submitted electronically through the paper submission system.

The copyright form is only needed after the paper has been accepted for publication in the proceedings.

Please note that papers must not exceed ten pages in length when typeset using the Procedia format.

Dates of deadlines for draft paper submission (full paper), notification of acceptance, deadline for camera-ready paper submission and registration may be found in the Important Dates section of this Web site.

Papers must be based on unpublished original work and must be submitted to ICCS only.

After the conference, selected papers will be invited for a special issue of the Journal of Computational Science.

Note: The paper submission system will open on September 15.