"Enchilada" is back on the menu

Stanislav Babak

Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, Potsdam-Golm, 14476, Germany

E-mail: `stas.babak@aei.mpg.de`

Abstract. In this short proceedings we describe the new pipeline for generating simulated LISA data. The pipeline relies on the catalogues of sources which represent the realisation of the Universe in gravitational waves (GWs) for several astrophysical models. We have extended the set of GW models which could be used to simulate the GW signals. Finally we have adopted hdf5 as the main format for the storage the data and parameters.

1. Past Mock LISA Data Challenges

Mock LISA Data Challenge is an effort started in January 2006 by a group of people (task-force) [1, 2]. The main idea behind this effort was to develop the common framework for the LISA data analysis. This includes defining LISA's orbit, noise budget, and set of models for simulating GW signals. Such a common ground simplify the comparison between different data analysis techniques, which are applied to the *same* data. We have used several simulators to apply the LISA response function [3, 4, 5]. The simulated data were advertised to scientific community and were available for download to whoever wanted to take part in the challenge. Each challenge was not a competition, but rather a way to foster the development and improvement of the LISA data analysis methods. There were two data sets in each challenge: (i) open, where all details about the data (parameters of the noise and each GW source) was known, (ii) blind challenge where only partial information (ranges) about number of sources and their parameters was disclosed.

Challenges were issued roughly once per year and overall were increasing in complexity. A short summary of the challenges is given in the table 1 and we will give details of each challenge a bit later. The participation has varied from challenge to challenge but in total we had 70 participants from 25 institutions. Before describing the challenges we should recall the main GW sources of LISA and their basic features. We refer reader to [6] for a more detailed description of LISA sources and science return.

The strongest expected GW signal will come from the coalescing massive black hole (MBH) binaries. Those are BHs residing in the nuclei of (we believe) all galaxies, they gain mass through the gas accretion and through minor/major mergers with other MBHs. The mergers of MBH is the result of galactic collisions, two MBHs interact with the surrounding gas and stellar environment and form a close binary where the gravitational radiation is efficient to drive them to a merger. The notion of minor or major merger refers to the mass ratio of coalescing MBHs, where the major corresponds to two MBHs of comparable mass. LISA will be sensitive to the MBH binaries in a wide range of masses from $10^4 M_{\odot}$ to few $10^7 M_{\odot}$. The expected event rate is quite uncertain from astrophysical point of view and also depends on the final configuration of

LISA, but we could expect between few and few hundreds events per year. We refer to [7] and references therein for more details. The expected signal-to-noise ratio (SNR) for those events varies in a wide range between 10 (we use as the detection threshold) and ~ 1000. The GW signal could be conventionally split in three parts: (i) Inspiral, this is low-frequency part of the signal, where the two bodies spiral around each other and adiabatically approach each other due to loss of energy and angular momentum by GWs, this is the longest part of the signal and could be observed (typically) for several months. This part of signal could be described analytically [8]. (ii) Merger, this part is of relatively short duration, two MBHs approach each other and form a single highly deformed object, this is accompanied by a huge release of energy in GWs and will be seen by LISA throughout the Universe (iii) Ringdown, this is high frequency part of the signal, where the deformed remnant MBH releases its excitations through the gravitational radiation, which spectrum is the superposition of exponentially damped quasi-normal modes of a MBH. This part of the signal will dominated the SNR for the heavy systems with the total mass $\gtrsim 5 \times 10^6 M_\odot$. The BHs are expected to be spinning where the magnitude of the spins and their orientation depend on the environment and the past history (for example, persistent accretion disk significantly spins up BHs). The spins misaligned with the orbital angular momentum cause precession of the orbital plane which encoded in the GW signal.

The most interesting GW in the LISA band is the so-called extreme mass ratio inspiral (EMRI). We expect that the MBHs in quiescent galaxies is surrounded by a cusp of compact objects (neutron star, solar mass BHs) which have segregated to the nuclei by dynamical friction mechanism [9]. As a result of N-body interaction one of such compact object could be deflected to the orbit passing near a MBH and being captured, then it slowly spirals toward the central MBH until the final plunge. The parameters of the orbit and the event rate should provide us with the unique information about central parsec region around MBH, which cannot be obtained by other means. The expected event rate varies from few per year to almost a thousand and SNR lies between 20 (detection threshold) and up to few hundred. The signal consists of harmonics of three non-commensurate (in general) orbital frequencies slowly changing in time as a small compact object spirals toward the central MBH. The signl spends in LISA band $10^4 - 10^6$ cycles and lasts up to few years. Fitting the signal's phase allows ultra-precise measurement of the source' parameters and this also makes it difficult to model such a GW signal. We refer to [10] for detailed overview of EMRIs (and references therein).

The most numerous source of GWs is our own Galaxy Milky Way. Gravitational radiation form $\sim 10^7 - 10^8$ white dwarf binaries on the tight orbits spreads across the LISA band. Majority of these GW signals will not be detectable individually and form a stochastic astrophysical foreground signal below few mHz. At high frequencies the density of those signals is low enough to resolve and subtract each signal. We will refer to residual Galactic GW noise signal (after subtraction) as "reduced Galactic signal". Each GW signal coming from the white dwarf binary is almost monochromatic with slight drift in frequency due to GW emission and/or due to mass transfer in the interacting binaries. The signals are always present in the data and stay in band. From the electromagnetic observations we know a dozen of Galactic binary systems which fall into a LISA band and will be detectable in GWs, those binaries are called *verification* binaries. Those a guaranteed sources with known sky position and GW frequency, and could be used for monitoring the detector's performance.

The three GW signals from binary black holes detected by LIGO have opened the era of gravitational wave astronomy [11]. The biggest surprise was the high mass of detected BHs ($\sim 30 M_\odot$), and those binaries at their early stages of inspiral (5-10 years before the merger) should emit a GW signal detectable by LISA [12]. Those signals will be detectable around/above 10 mHz, and the population of binary BHs might create a detectable stochastic GW signal.

Besides the GW signal from coalescing binary systems described above we might be able to detect a stochastic GW signal from very energetic processes in early Universe [13]. Such a GW signal is expected to be isotropic and its detection is based on the cross-correlation of independent data streams constructed out of individual measurements from each link (laser signal connecting two spacecrafts). The stochastic signal from the astrophysical populations described above should have a specific power-law spectral index, which should allow to distinguish it from the stochastic signal of cosmological origin.

Finally, LISA might detect something exotic like GW burst signal coming from cusps or kinks formed on the cosmic strings (see [13] and references therein). those signals have a very specific shape in time and frequency domain which allows to search for them using matched filtering techniques.

Now we are in the position to describe the past MLDCs summarized in the table 1. The very first channel was the easiest, it contained several independent data sets (i) data set with few Galactic binaries (ii) data set with the verification binaries (iii) data set with few dozen of Galactic binaries (iv) data set with a single GW signal from non-spinning MBH binary with SNR$>$ 100. The second challenge was quite a big jump ahead, it contained again several independent data sets: (i) about 20 mln. monochromatic signals from Galactic binaries (ii) 4-6 (exact number was not disclosed) GW signals from non-spinning MBH binaries placed on top of the reduced Galactic signal (iii) 5 data sets each with a strong (SNR$>$ 50) GW signal from EMRIs. The next challenge was a repetition of the challenge 1 to allow new groups to join in and to allow further investigations and improvements in the data analysis algorithms. The MLDC3 had again many data sets each containing: (i) 60 mln. GW signals with frequency drift from Galactic binaries (ii) 4-6 GW signals from spinning (precessing) MBH binaries together with the reduces Galactic signal (iii) single data set with 5 weak (SNR$\approx$ 20) EMRIs (iv) data set with GW bursts from cusps on cosmic strings (v) data set with the isotropic stochastic GW signal. For the population of Galactic binaries we have used catalogue provided by G. Nelemans [14, 15].

Table 1. Summary of past mock LISA data challenges

Sources	MLDC1	MLDC2	MLDC1B	MLDC3
Gal. bin	Verification, few dozen	Galaxy, no freq. evolution	Verification, few dozen	Galaxy 6×10^7
MBH bin.	One, no spin	4-6, no spin, with red. Gal.	One, no spin	4-6 precess., with red. Gal
EMRIs	-	Isolated, strong	Isolated strong	Isolated, weak
Bursts	-	-	-	Cosmic strings
Stochastic	-	-	-	Isotropic, strong

2. Recipe for “Enchilada”

From the table 1 and from the previous description it is clear that the current data analysis methods have mainly concentrated on the detection and characterization of GW signals of the same type in the data sets where only few signals are present. Of course the analysis of Galactic binaries is an exception, there we meet the problem of the *signal confusion*, when multiple signals overlap in time and in frequency, but those signals are of the same type. Several data analysis methods were suggested to tackle problem of detecting Galactic binaries (see [16, 17] and references therein).

The “Enchilada” was suggested for the first time for the challenge 2 but it was premature, later it was a theme for the never-finished challenge 4. The “Enchilada” is a nick name (suggested by N. Cornish) for a challenge where the signals of the same and different kind are simultaneously present in the data. The Enchilada addresses the main problem of LISA’ data analysis: confusion problem, where we need to detect and characterize (correctly) the maximum possible number of

GW signals. Besides the problem of detecting Galactic binaries, the only mix of sources was the GWs from MBH binaries plus reduced Galactic signal, but the reduced Galactic signal played role of the noise source.

The challenge 4 (which was never completed) was meant as a mild Enchilada problem, and here we want resurrect it but with improved "recipe". So what do we need to cook up the Enchilada data set.

- We need to simulate the Universe as it can be seen by LISA in GWs. We need astrophysical models which tell us how many signals we expect to have and distribution of parameters.
- We need reliable and fast to generate models of GW signals from anticipated sources.
- We need to simulate the instrument: mission configuration, noise level, instrumental and environmental artifacts.

We will walk through each of these items. The first item is about building catalogues of sources based on the current astrophysical models. The uncertainties in our knowledge/models will be reflected in several catalogues. For merging MBH binaries we used the models described in [7]. We also have catalogues for EMRIs [18] which was built on top of the MBH population [19, 20]. The novelty there (brought by LIGO) is that we have added heavy solar-origin BHs ($\sim 30 M_{\odot}$). We have also improved catalogue of Galactic sources. The idea is to use the sources selected from these catalogues for the simulated data sets. The flexibility of a current data generating pipeline allows us to use a whole realization of a given model or just a selected number of sources.

The second bullet point is about simulating the GW signals. There are two key points which must be satisfied by a model: (i) it should be effectual representation of the GW signal (ii) it must be fast to generate. The former implies that model might not be necessarily very faithful, of course it is desirable. The later is required for the efficient exploration of the data analysis methods, and it wins over faithful but slow models. For coalescing MBH binaries we have extended the pool of models by including IMR models (models including all three parts of the signal inspiral-merger-ringdown), whereas previously we have used only inspiral part of the signal. We have added two models for precessing MBHs the EOB-based time domain model [21] and the phenomenological frequency-domain model [22]. The time domain model is convenient for the simulating the data, whereas the frequency domain model is more suitable for the data analysis purposes. This is an important extension, as the merger-ringdown is a very important part from the point of view of detection and parameter estimation. For EMRIs we use the model suggested in [23], however this model is notoriously unfaithful, so we have added also the model suggested in [24].

The third bullet point is actually very big. When we say GW signal model we usually mean two GW polarizations $(+, \times)$, however LISA has a non-trivial response function (see, for example, [4]). What is used for data analysis are two (noise) independent TDI streams, called A, E, which are result of applying the response/transfer function to GW polarizations. The TDI stands for time delay interferometry, which is a linear combination (with time delays) of individual measurements which cancel laser frequency noise (the highest noise) [25]. We use LISACode [5] to apply the response and derive the A, E combinations for each GW signal. The LISACode has a flexibility to vary the LISA configuration (like arm-length, mission life time, size of collecting telescope, power of laser) as well as the basic noise budget. This is an important property which we want to have before the final LISA configuration is stable. We also have now LISA Pathfinder data [26] which exhibits fantastic performance. The measured data is however shows occasionally non-stationary features. We want to explore the effect of these noise artifacts on the data analysis algorithms which were developed with assumption of the Gaussian noise. For the Enchilada project we start with the Gaussian noise of a given color (derived from LISA Pathfinder measurements projected to LISA) and LISACode will be used to produce such a

f_ini	0.000427690264488	Hz
spin2x	-0.0891075135401	dimensionless
mass1	280533.8732	solar mass
mass2	39644.2998	solar mass
spin2z	0.576370782859	dimensionless
spin2y	-0.0366724501106	dimensionless
longitude	6.17673316892	rad
polarization	2.63132015174	rad
spin1x	0.181940330537	dimensionless
spin1y	0.132187387695	dimensionless
Lum_Distance	19299.8833863	Mpc
spin1z	0.762715833754	dimensionless
latitude	-0.0116491833204	rad
phi_ref	5.99308322852	rad
model	SEOBNRv3	approximant
cadence	3.0	sec
inclination	1.5333	rad

Figure 1. Example of the binary MBH parameter table using the small viewer supplied with the data.

noise data stream. Later we intend to extend this project by including the simulated realistic noise model.

3. Implementation

In this section we briefly discuss the implementation. The biggest difference with the previous MLDC pipeline is that we move to the new format, we intend to use hdf5 to store the parameters and the data. Those files will also contain some minimum metadata information (author, date, software version) which should be sufficient for reproducing the stored data. The hdf5 library has many APIs with different programming languages and we will provide a read/write interface with python and C/C++. There are standard hdf5 viewers which can easily parse the files and present the parameters as tables. We also provide a small viewer which allows to visualize the source parameters, an example for the a single MBH binary is given in the figure 1.

Previously we have used xml-format to store the metadata and links to the binaries with the simulate data [27], and indeed it is convenient format in some cases. We still can store the old xml files as a string in hdf5 files.

We have successfully produce several test data streams using new waveform models. We show example of two GW signals from merging MBH binaries in the figure 2. The red system (short one) is seen almost edge-on, and parameters of that system is given in the table of the figure 1, the clearly seen precession is caused by the misalignment of spins and the orbital angular momentum. For the long blue signal we have only very slight misalignment of spins and precession is not prominent.

In figure 3 we have put together GW signals from Galactic binaries, two MBH binaries (given in the figure 2) and one signal from EMRI. The left panel shows the simulated data in time domain and the right panel is power spectral density of the data. In particular we plot here X-TDI Michelson data combination. One can see in the time series that all signals are instantaneously below the noise level, even though they are detectable. The frequency domain plot shows the frequency range for each signal. Note that the merger part of MBHs signal and EMRIs could reach quite high frequency and require a small cadence (3-5 sec). This potentially could restrict the duration of the challenge data sets. This data set was obtained for the LISA

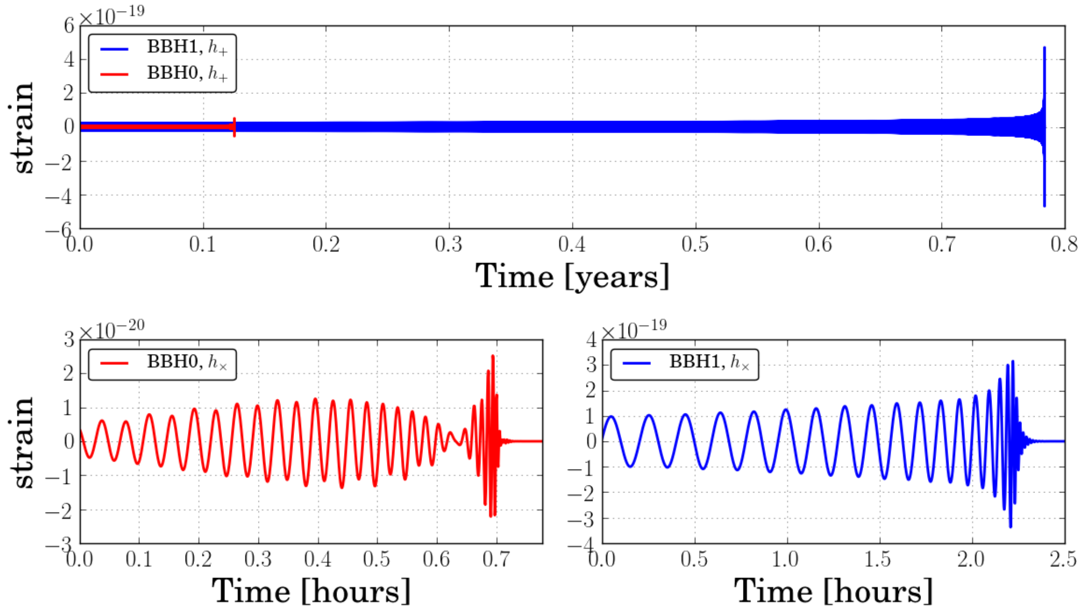

Figure 2. Two GW signals from MBH binaries. The top panel shows h_+ and lower bottom is a zoom in around the merger for othe polarization $h_\times$

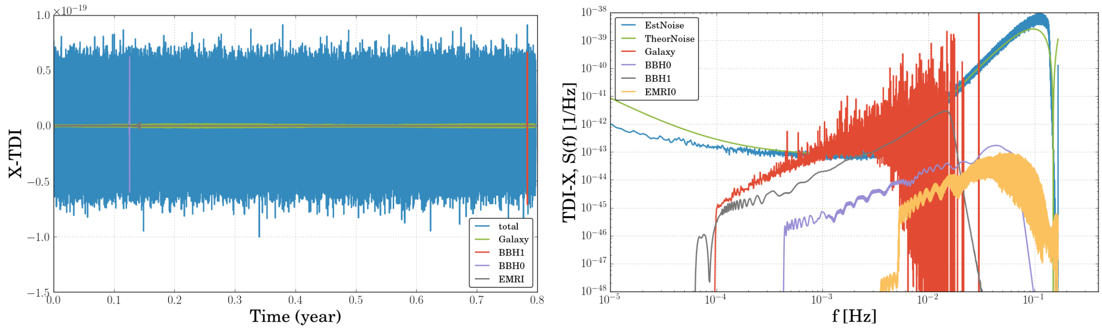

Figure 3. Example of the simulated data in time domain (left panel) and in frequency domain (power spectra density of the data in the right panel). The data contains simulated noise, 2 MBH binaries, 1 EMRI signal and a whole catalogue of the Galactic white dwarf binarie ($\sim 6 \times 10^7$ sources).

configuration with 1 mln. km armlength.

4. Discussion

We have pipeline for data set production, but we are not yet clear about the format of the new mock data challenges. There will be a core group which will analyze the data as it is required for the LISA project. With LISA on the clear roadmap, the funding agencies might be willing to support LISA data analysis related projects, and we want to make the challenges globally available. Another important question is about complexity of the data, we do not want to make a fake impression that the problem is solved by analysing the simplified data set, and we do not want to pass the wrong message that we cannot handle complexity of the LISA data. Probably we will make few initial trials "in-house" trying to analyze few data sets starting with a very simple one. In the mean time we might repeat single-source type challenge (but using the new

GW models) to dust off the tools used in the challenge 3, and get momentum for the "Enchilada" problem.

The data will be issued again in two formats: open, where all details about the data (noise, number of GW signals and their parameters) are known, and, closed data set which is blind data challenge, where only minimum information will be given to participants. We will make available all GW models (software) which were used for production of the challenge data sets, however some additional work (code optimization) might be required for the data analysis purposes.

Finally, we would like to say few words about the data analysis status. During the past MLDC several very promising methods were developed. Among the frequentist methods which aim at maximimization of the likelihood and getting the maximum likelihood estimators for parameters, we should mention the grid-based methods, Genetic Algorithm and particle swarm optimization. Bayesian methods were also used to analyse the data, those are various modifications of the Markov chain monte-carlo technique (simulated annealing, parallel tempering, etc.), and nested sampling (MultiNest). Note that there is a large community analysing the LIGO data, and some aspects of the LIGO data analysis are common with the LISA, so we can port the tools and expertise developed in LIGO to the MLDC. In addition to LIGO tools and methods, the problem of GW analysis of the Pulsar Timing Array (PTA) data is very similar to the Galactic GW foreground problem in LISA. One of the main source in the PTA band is a population of broad massive BH binaries in the local Universe. The GW signals which they emit are monochromatic over the observation time and they form a stochastic signal at low frequencies. Some GW signals (from nearby sources) could stand above of that GW background. At high frequencies the density of sources is dropping down and the GW signals should be detectable individually (when the sensitivity reaches the detectability level).

The new Mock LISA data challenge comes up this year. The most challenging problem which we will focus on is the signal confusion, also referred as Enchilada. We will used catalogues of sources based on the astrophysical models and we apply the new (extended) models for the GW signal simulation. Finally we adopt the new (hdf5) format for storing and distributing the data and associated metadata.

References

[1] Arnaud K A *et al.* (Mock LISA Data Challenge Task Force) 2006 *AIP Conf. Proc.* **873** 625–632 [,625(2006)] (*Preprint* `gr-qc/0609106`)
[2] Arnaud K A *et al.* 2006 *AIP Conf. Proc.* **873** 619–624 [,619(2006)] (*Preprint* `gr-qc/0609105`)
[3] Vallisneri M 2005 *Phys. Rev.* **D71** 022001 (*Preprint* `gr-qc/0407102`)
[4] Rubbo L J, Cornish N J and Poujade O 2004 *Phys.Rev.* **D69** 082003 (*Preprint* `gr-qc/0311069`)
[5] Petiteau A, Auger G, Halloin H, Jeannin O, Plagnol E, Pireaux S, Regimbau T and Vinet J Y 2008 *Phys. Rev.* **D77** 023002 (*Preprint* `0802.2023`)
[6] Amaro-Seoane P, Aoudia S, Babak S, Binetruy P, Berti E *et al.* 2013 *GW Notes* **6** 4–110 (*Preprint* `1201.3621`)
[7] Klein A *et al.* 2016 *Phys. Rev.* **D93** 024003 (*Preprint* `1511.05581`)
[8] Blanchet L 2014 *Living Reviews in Relativity* **17** URL `http://www.livingreviews.org/lrr-2014-2`
[9] Alexander T and Hopman C 2009 *Astrophys. J.* **697** 1861–1869 (*Preprint* `0808.3150`)
[10] Babak S, Gair J R and Cole R H 2015 *Fund. Theor. Phys.* **179** 783–812 (*Preprint* `1411.5253`)
[11] Abbott B P *et al.* (Virgo, LIGO Scientific) 2016 *Phys. Rev.* **X6** 041015 (*Preprint* `1606.04856`)
[12] Sesana A 2016 *Phys. Rev. Lett.* **116** 231102 (*Preprint* `1602.06951`)
[13] Babak S, Gair J R, Petiteau A and Sesana A 2011 *Class.Quant.Grav.* **28** 114001 (*Preprint* `1011.2062`)
[14] Nelemans G, Yungelson L R and Portegies Zwart S F 2001 *Astron. Astrophys.* **375** 890–898 (*Preprint* `astro-ph/0105221`)
[15] Nissanke S, Vallisneri M, Nelemans G and Prince T A 2012 *Astrophys. J.* **758** 131 (*Preprint* `1201.4613`)
[16] Babak S *et al.* (Mock LISA Data Challenge Task Force) 2008 *Class. Quant. Grav.* **25** 114037 (*Preprint* `0711.2667`)
[17] Babak S *et al.* (Mock LISA Data Challenge Task Force) 2010 *Class. Quant. Grav.* **27** 084009 (*Preprint* `0912.0548`)
[18] 2017 In preparation
[19] Sesana A, Barausse E, Dotti M and Rossi E M 2014 *Astrophys. J.* **794** 104 (*Preprint* `1402.7088`)

[20] Antonini F, Barausse E and Silk J 2015 *Astrophys. J.* **812** 72 (*Preprint* 1506.02050)
[21] Babak S, Taracchini A and Buonanno A 2017 *Phys. Rev.* **D95** 024010 (*Preprint* 1607.05661)
[22] Hannam M, Schmidt P, Bohé A, Haegel L, Husa S, Ohme F, Pratten G and Pürrer M 2014 *Phys. Rev. Lett.* **113** 151101 (*Preprint* 1308.3271)
[23] Barack L and Cutler C 2004 *Phys.Rev.* **D69** 082005 (*Preprint* gr-qc/0310125)
[24] Chua A J K and Gair J R 2015 *Class. Quant. Grav.* **32** 232002 (*Preprint* 1510.06245)
[25] Tinto M and Dhurandhar S V 2005 *Living Reviews in Relativity* **8** URL http://www.livingreviews.org/lrr-2005-4
[26] Armano M *et al.* 2016 *Phys. Rev. Lett.* **116** 231101
[27] Vallisneri M and Babak S 2008 *Computing in Science & Engineering* **10** 80–87 (*Preprint* http://aip.scitation.org/doi/pdf/10.1109/MCSE.2008.20) URL http://aip.scitation.org/doi/abs/10.1109/MCSE.2008.20