Job Demand Models for Optical Grid Research

Konstantinos Christodoulopoulos[1], Emmanouel Varvarigos[1],
Chris Develder[2], Marc De Leenheer[2], Bart Dhoedt[2]

1: Dept. of Computer Engineering and Informatics, and
Research Academic Computer Technology Institute
University of Patras, 26500, Patras, Greece
2: Dept. of Information Technology (INTEC), Ghent University – IBBT,
G. Crommenlaan 8 bus 201, 9050 Gent, Belgium
{kchristodou, manos}@ceid.upatras.gr
{chris.develder, marc.deleenheer, bart.dhoedt}@intec.ugent.be

This paper presents results from the IST Phosphorus project that studies and implements an optical Grid test-bed. A significant part of this project addresses scheduling and routing algorithms and dimensioning problems of optical grids. Given the high costs involved in setting up actual hardware implementations, simulations are a viable alternative. In this paper we present an initial study which proposes models that reflect real-world grid application traffic characteristics, appropriate for simulation purposes. We detail several such models and the corresponding process to extract the model parameters from real grid log traces, and verify that synthetically generated jobs provide a realistic approximation of the real-world grid job submission process.

Keywords: Optical Grids, Job demand models, Profiling, Expextation Maximization algorithm, Probabilistic modeling

1. Introduction

Today, the need of network systems for storage and computing services for scientific and business communities are often answered by relatively isolated islands, known as clusters. Migration to truly distributed and integrated applications requires optimization and (re)design of the underlying network technology. This is exactly what Grid networks promise to offer: a platform for cost and resource efficient delivery of network services to execute tasks with high data rates, processing power and data storage requirements, between geographically distributed users. Realization of that promise requires integration of Grid logic into the network layers. Given the high data rates involved, optical networks offer an undeniable potential for the Grid. An answer to the demand for fast and dynamic network connections could lie in the (relatively) new switching concepts such as Optical Packet Switching (OPS) and Optical Burst Switching (OBS) [1].

Delivering the Grid promise implies answering a series of fundamental questions [2]: (re)design the architecture of a flexible optical layer, development of the

I. Tomkos et al. (Eds.): ONDM 2007, LNCS 4534, pp 127-136, 2007.

necessary design techniques for e.g. dimensioning, algorithms for routing and control offering both QoS and resilience [3] guarantees. It is this—to a large extent unexplored [4]—area of fundamental research that is the subject of the supporting studies within the Phosphorus project.

Many of the answers to these research questions are addressed through simulations. A necessary prerequisite to obtain useful results is an adequate model of the traffic (i.e. jobs) that will be submitted to the Grid. Although, a great deal of work has appeared in literature on job characterization and modeling for single parallel supercomputers [5], similar work in the area of (optical) Grids is quite limited. Medernach [6] analyzed the workload of an LCG/EGEE cluster, proposing a 2-dimensional Markov chain for modeling single user behavior in a Grid. Li et al. [7] used the LCG Real Time Monitor to collect data from the global EGEE Grid, and proposed models at three different levels: Grids, Virtual Organizations and regions. They conclude that Markov Modulated Poisson Processes (MMPP) with sufficient number of states can reflect the real world job arrival processes. In the work presented here, we introduce a new model, referred to as the Pareto-Exponential model and compare it with the previously proposed model. We show that, despite its more compact parameter set, the Pareto-Exponential model is a valid alternative. Finally, we also propose a model for the job execution times that is based on the hyper-exponential distribution.

We start in Section 2 by briefly outlining the Phosphorus project and how the job demand modeling work fits in the whole concept. In Section 3 we introduce the candidate traffic models considered, and discuss how we fitted them to real world traffic traces at different aggregation levels in Section 4. Finally, Section 5 summarizes our conclusions.

2. The Phosphorus project

As indicated in the introduction, a new generation of applications is emerging, coupling data and high-end computing resources distributed on a global scale. These impose requirements such as determinism (e.g. guaranteed QoS), shared data spaces, large data transfers, that are often achievable only through dedicated optical bandwidth. High capacity optical networking can satisfy bandwidth and latency requirements, but software tools and frameworks for end-to-end, on-demand provisioning of network services need to be developed in coordination with other resources (CPU and storage) and need to span multiple administrative and network technology domains.

In response to the above requirements, the European IST project Phosphorus will address some of the key technical challenges to enable on-demand end-to-end network services across multiple domains. The Phosphorus network concept and testbed will make applications aware of the Grid environment, i.e. the state and capabilities of both computational and network resources. Based on this information, it is possible to make dynamic, adaptive and optimized use of heterogeneous network infrastructures connecting various high-end resources. The testbed will involve European NRNs and national testbeds, as well as international resources (GÉANT2, Internet2, Canarie, Cross Border Dark Fibre infrastructures and GLIF virtual facility). A set of highly demanding applications will be adapted to prove the concept.

In the Work Package 5, "Supporting Studies", architectural and algorithmic questions will be addressed. These include research in the area of job routing and scheduling algorithms (decide where to execute a given job and how to reach that destination, referred to as the anycast routing problem [8]), examine techniques that jointly reserve computation and communication resources, and compare packet versus circuit switching technologies.

3. Job Demand Models

To objectively evaluate the performance of e.g. job scheduling and routing algorithms, it is desirable that the job submission model accurately reflects the characteristics of real world grid jobs. When simulation is used, an analytical model is preferred over actual job traces, since this approach allows the different job parameters (e.g. average load) to easily be adjusted. In this section we present such analytical models for the job arrival/processing times.

The classical Poisson process model, in which the inter-arrival times are exponentially distributed, forms the basis for most of these more advanced models.

3.1. Non-Homogeneous Poisson Process (NHPP)

As can be easily intuitively accepted, the job arrival rate can exhibit time dependent behavior, especially on e.g. national scales. On a daily scale, day/night differences can be observed, and also the difference between week days and weekends may be visible. Hence, the arrival rate of a Poisson process can be considered to be a function of time λ(t), leading to a non-homogeneous Poisson process. More specifically, the number of arrivals N(t) in the interval [0, t] follows the distribution shown in (1).

$$\Pr[N(t)=n]=e^{-m(t)}\frac{(m(t))^n}{n!}, n\geq 0 \quad \text{and} \quad m(t)=\int_0^t \lambda(s)ds \tag{1}$$

In fitting this model to real life job traces (described in Section 4), we will consider the job arrival rate λ(t) to be a stepwise function. In this case, the job generation model can be considered as a state process, where the system evolves from one state to the next while maintaining a fixed arrival rate λ_i that depends only on the state *i*.

3.2. Phase-type process

An *m*-phase type distribution represents a random variable (in our case e.g. job inter-arrival times) whose values are the transition times until absorption of a continuous-time Markov chain with *m* transient states and one absorbing state. In general, any inter-arrival process can be approximated by a phase-type distribution provided enough states are introduced. A special case of the general phase-type distribution is the *hyperexponential* distribution (HE), which has two or more non-identical phases that occur in parallel (i.e. each of the phases only has a non-zero probability to transit to the absorbing state). The probability density function (pdf) of a hyperexponentially distributed variable X is given in (2). This corresponds to the weighted sum of m exponentially distributed random variables Y_i (with average $1/\lambda_i$).

$$f_X(x) = \sum_{i=1}^{m} p_i f_{Y_i}(y) \text{ where } \sum_{i=1}^{m} p_i = 1 \qquad (2)$$

3.3. Markov Modulated Poisson Process (MMPP)

The Markov modulated Poisson process (MMPP) is a doubly stochastic Poisson process [9], characterized as a (finite state) continuous time Markov chain with m states. Each state i is a Poisson process (arrival rate λ_i) in itself, and state transitions are defined by a state transition matrix $\boldsymbol{Q}$. Thus, the system is fully defined by a matrix $\boldsymbol{Q}$, as defined in (3), and a vector $\boldsymbol{\Lambda} = [\lambda_1 \lambda_2 \ldots \lambda_m]$.

$$\boldsymbol{Q} = \begin{bmatrix} -\sigma_1 & -\sigma_{1,2} & \cdots & -\sigma_{1,m} \\ -\sigma_{2,1} & -\sigma_2 & \cdots & -\sigma_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ -\sigma_{m,1} & -\sigma_{m,2} & \cdots & -\sigma_m \end{bmatrix} \text{ where } \sigma_i = \sum_{j=1, i\neq j}^{m} \sigma_{i,j} \qquad (3)$$

3.4. Pareto-exponential model (PE)

In the Pareto-exponential model busy periods, in which jobs arrive, succeed each other. Each busy period has an exponentially distributed duration (with mean of $1/\mu$ seconds), and within a busy period jobs arrive according to a Poisson process (at a rate of λ jobs/second). The times between the start times of a busy period are distributed following a truncated Pareto distribution with shape parameter α, minimum value X_{min} and maximum value X_{max}. An intuitive interpretation of the busy periods can be that these correspond with job submissions from a particular virtual organization (VO) participating in the Grid.

4. Real life measurements

To validate the suitability of the various models for the job arrival process and their execution times, we have collected traces from operational Grid environments. We then fitted the aforementioned models to the traces, and used the parameter values found to drive a simulator generating the IATs according to the respective models with the parameter values found by the fitting algorithm.

4.1 Measured infrastructure

Since the Phosphorus test-bed is still under construction, we gathered traces on the LCG/EGEE infrastructure [10]. The Enabling Grids for E-sciencE (EGEE) project offers an always-on Grid computing infrastructure, geographically distributed across the globe. The worldwide LHC Computing Grid project (LCG) was created to offer computing infrastructure processing and analyzing the data of the Large Hadron Collider (LHC) experiments at CERN. The LCG and EGEE projects share a large part of their well established infrastructure; hence we refer to it as the LCG/EGEE infrastructure. Currently, it comprises 207 cluster sites from 48 countries. In the observation period, we recorded the presence of 39,697 CPUs (of which on average 31,228 were active) and 5 Petabytes of storage.

The job lifetime comprises various phases, as illustrated in Fig. 1. Users submitting jobs are part of a VO (Virtual Organisation), which is a dynamic collection of individuals and institutions sharing the same permissions etc. In order to submit jobs to the Grid, a user has to log in to a user interface (UI) and provide the job specification in a JDL (job description language) format. This job submission is then forwarded to the corresponding Resource Broker (RB), which will schedule the job for execution taking into account information provided by the JDL as well as information service (e.g., the VO, global traffic load information). The job, wrapped in an input sandbox, is eventually sent to a Computing Element (CE) at a particular site, where a local resource management system will assign it to a Worker Node (WN).

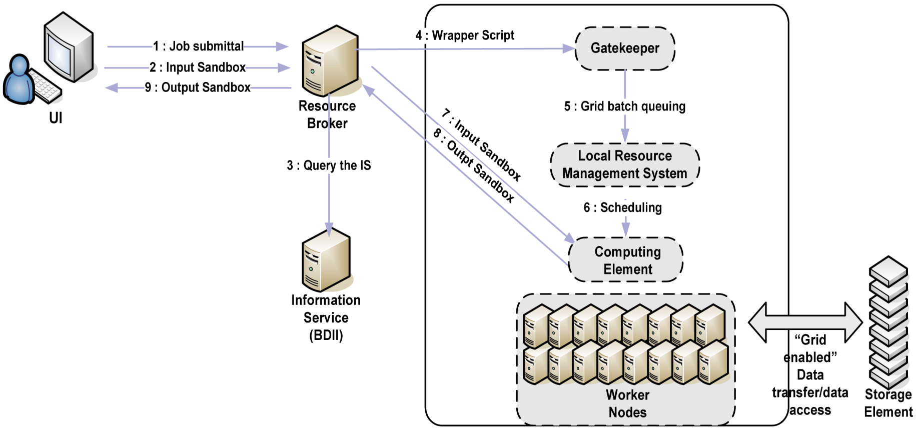

Fig. 1. Job flow in the Grid environment.

As indicated, a job evolves through various states. Of particular interest to optical grid job modelers are (i) the job inter-arrival times (IATs), and (ii) the time spent in the "Running" state, which amounts to actual execution time of a job, including the I/O time. In the following we will establish suitable models by fitting them to the measured data. The subsequent subsection details the fitting methodology used.

4.2. Trace fitting methodology

In order to fit the distribution parameters to measured data samples, we use a maximum-likelihood estimation (MLE) technique, and specifically the Expectation-Maximisation (EM) algorithm. Further details can be found in [11].

4.3. Jobs at the Grid level

Using the LCG Real Time Monitor [11], we collect data for jobs submitted to all Resource Brokers (RBs) participating in the EGEE project. The RTM records the times at which user jobs are submitted, the way they are distributed to the sites, the times at which the jobs complete the different states of their processing, and finally depending on the successful or not execution it also presents the times of delivery of the execution outcome to the corresponding user. Of main interest here were the job IATs and running time. For this, we collected job arrival data during 1-31 Oct. 2006 (totaling 2,228,838 jobs).

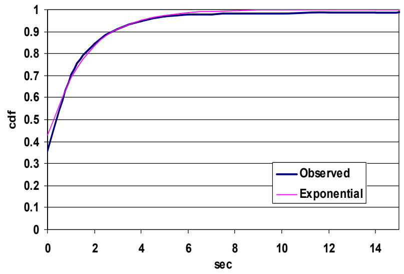

Fig. 2. Empirical cdf and Poisson process fit for IAT at the Grid level.

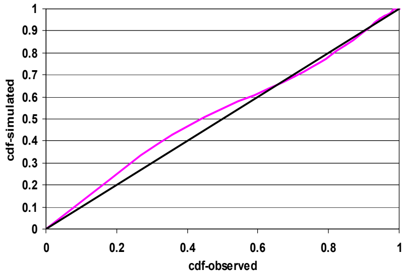

Fig. 3. Empirical vs Poisson P-P plot for IAT at the Grid level.

The empirical cumulative distribution function (cdf) is shown in Fig. 2. Given the observed standard deviation being close to the mean IAT, and the absence of a heavy tail, it is clear that a Poisson process can be a suitable model. Using MLE fitting, we found a Poisson process with mean IAT ($1/\lambda$) of 1.6077. Note that since our measurement data has a resolution of 1 second, we actually converted the Poisson process IATs by rounding them to the closest integer.

In Fig. 3, the probability-probability (P-P) plot (composed by pairing percentiles corresponding to the same value) is shown for the rounded Poisson model versus the empirically observed data. This plot being close to the line between (0,0) and (1,1) we may conclude the Poisson process model is adequate.

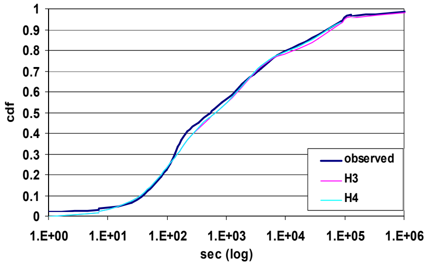

Fig. 4. Empirical cdf and HE fit for job execution time at the Grid level.

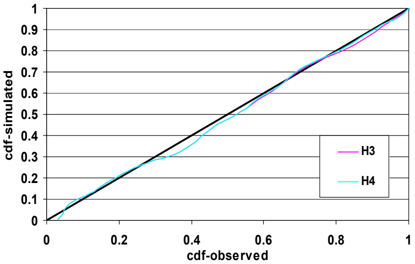

Fig. 5. Empirical vs HE P-P plot for job execution time at the Grid level.

- The WN execution times exhibit peaks at certain time values, reflected in a sharp rise in the cdf as depicted in Fig. 4. Thus, we resorted to fits with the hyper-exponential model, one with 3 phases (H3) and one with 4 phases (H4), based on the observation that the cdf curve exhibits 3-4 “steps”. We used the EMpht utility [12] implementing the EM algorithm described in [14].

To assess the quality of these fits, we again generated time values by simulation and obtained the cdfs (Fig. 4) and P-P plots (Fig. 5). Since the accuracy of H3 and H4 fits is similar, we can conclude that the 3-phase exponential process is sufficient to model the WN execution times at the Grid level.

4.4. Jobs at the Grid Site level

In addition to the Grid level traces, we also collected information for jobs submitted to individual Grid sites (computing elements in the scheme of in Fig. 1).

4.4.1. Kallisto site

The Kallisto node located in Patras is part of Hellasgrid and has been a production site since 1 Feb. 2006. It comprises 64 Intel Xeon CPUs at 3.4GHz and 2Gb RAM, of which 60 are actual Worker Nodes (the 4 other service nodes comprise the EGEE core servers), using g-Lite middleware and running scientific Linux v3.

For the measured job IATs, we considered the four different models discussed in Section 3:

- *Non-Homogeneous Poisson Process (NHPP):* We defined a stepwise function for λ(t), being constant over 1 hour intervals, with hourly values of λ obtained by averaging over all days in the observation period.
- *Hyper-Exponential Model:* We considered (i) a 2-phase (H2), and (ii) a 3-phase (3H) hyper-exponential model, based on the observation of 2-3 steps in the empirical cdf. Fitting was done using the EMpht software.
- *Markov Modulated Poisson Process (MMPP):* We considered (i) a 3-state (3MMPP), and (ii) a 4-state (4MMPP) MMPP.
- *Pareto-Exponential Model:* In this case, we chose a truncated Pareto distribution with X_{max}=10800 sec, since this was the (deterministic) periodicity of site

functional test. For the other parameters, we fitted: λ=18 arrivals/sec for busy periods, mean duration $1/\mu$=22.5 sec of the busy periods, a=0.48 and X_{min}=32 sec.

To evaluate the applicability of the models, we generated synthetic job traces using a simulator implementing the models, resulting in the cdf graphs of Fig. 6 and the corresponding P-P plots shown in Fig. 7.

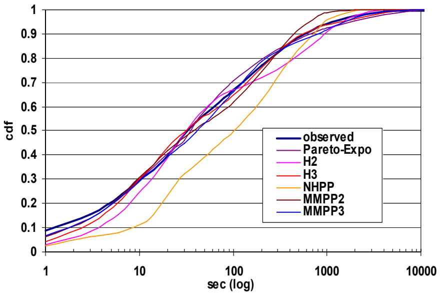

Fig. 6. Empirical cdf and various models fitting for IAT at Kallisto.

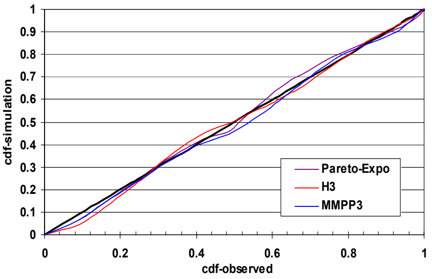

Fig. 7. Empirical vs various models P-P plot for IAT at Kallisto.

From the above graphs, we can conclude that conclude that the proposed Pareto-Exponential model generates traces that are very close according to the P-P plot to those observed in our cluster. H3 and 3MMPP models also simulate satisfactorily the job arrival process. However, the Pareto-Exponential model is simpler, more concise and more intuitive than the other proposed models, since it is based on a smaller number of parameters, and seems to correspond to actual VO behavior. With respect to the HE and MMPP models we observe, as expected, that the fit improves by increasing the number of phases (states for MMPP).

We have also computed the Hurst parameter for the four models (as a measure of long range dependency, aka self-similarity). Only the Pareto-exponential and the MMPP models experience long-range dependence (H=0.58 for PE, H=0.62 for 2MMPP and H=0.64 for 3MMPP with confidence levels higher than 99%). The value in the measured data amounted to H=0.68.

With respect to the job execution times (page limitations prevent us from showing the details), we observed similar behavior as on the Grid level, and fits using the 3H and 4H models proved to produce adequate results. On a quantitative level, we observed some differences which are due to the smaller number of VOs served by the Kallisto node compared to the complete EGEE (11 vs 75 VOs, with the most active VO in Kallisto –ATLAS VO- being the 3rd on a global scale).

4.4.2. BEGrid site

The BEGrid site located in Ghent is part of the Belnet grid initiative, originating in 2003. It comprises 41 dual Opteron (1.6 GHz) worker nodes, and another 15 dual dualcore Opteron (2 GHz) nodes, all having 4Gb RAM, and 5 service nodes, using g-Lite middleware and running scientific Linux v3.

In contrast to the Kallisto and global EGEE results, we found a far less smooth, more step-wise empirical cdf of the job IAT for the BEGrid measurements. We found a particularly steep increase, corresponding to a peak in the probability density function (pdf) around 8-15 seconds. The reason for this is that a large group of BEgrid users commits their jobs using scripting, with script submission overhead resulting in a job IAT of the order of 10s. Obviously, neither the exponential distribution nor the more complex functions succeed in reproducing this abrupt cdf. Hence the poor P-P plots.

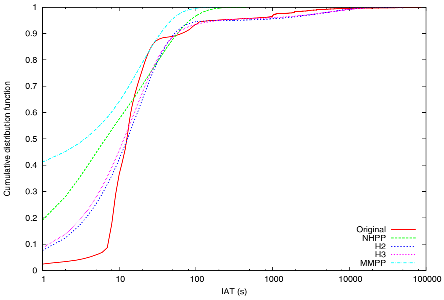

Fig. 8. Empirical cdf and various models fit for IAT at BEGrid site.

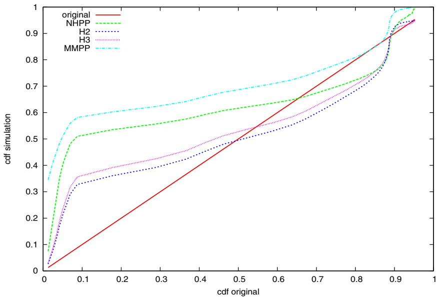

Fig. 9. Empirical vs various models P-P plot for IAT at BEGrid site.

With respect to the job execution times, the observed cdf (omitted because of space limitations) was of a similar slightly step-wise shape as the EGEE and Kallisto measurement data. Again, we found 3- and 4-phase hyper-exponential models to generate the most satisfactory results.

5. Conclusions

Due to the high equipment cost involved in the research of optical grids if actual hardware were to be used, simulation techniques are often put forward as a viable alternative. To warrant accurate and useful results, it is important that a realistic grid job load is used as input for the simulation. To this end, we presented analytical job models, and the methodology to extract model parameters from actual grid log traces. This approach guarantees a very flexible, analytical job submission model, yet providing a very realistic approximation of the real life grid job submission pattern.

Using real life measurement data, gathered at different aggregation levels in a Grid environment (local site vs global Grid), we judged the usefulness of various models fitted to that data. This was achieved by implementing the models in simulation software. From this study, we concluded that:

- Job inter-arrival times on the observed Grid level can be successfully modeled by a Poisson process, but on the Grid site level (eg. Kallisto traces) the long range dependency needs to be taken into account and HP, MMPP or Pareto-Exponential models need to be used.
- For the job execution times, we achieved the most satisfactory results with a (3 phase) hyper-exponential process.

Acknowledgements

The work presented in this paper was supported by the EU through the IST Project Phosphorus (www.ist-phosphorus.eu) and e-phton/ONe+ (www.e-photon-one.org). C. Develder thanks the Research Foundation – Flanders (FWO – Vlaanderen) for his post-doctoral fellowship. M. De Leenheer thanks the IWT for financial support through his Ph. D. grant.

References

1. C. Develder, et al., “Delivering the Grid Promise with Optical Burst Switching” (Invited), Proc. Int. Workshop on Optical Burst/Packet Switching (WOBPS East 2006), at the Joint Int. Conf. on Optical INternet and Next Generation Networks (COIN-NGN 2006), Jeju, South Korea, Jul 2006.
2. D. Simeonidou, et al., "Dynamic Optical Network Architectures and Technologies for Existing and Emerging Grid Services", IEEE Journal of Lightwave Technology, 23(10):3347-3357, Oct 2005.
3. J. Zhang, and B. Mukherjee, “A review of fault management in WDM mesh networks: basic concepts and research challenges”, IEEE Network, 18(2):41-48, Feb 2004.
4. K.H. (Kane) Kim, “Wide-Area Real-Time Distributed Computing in a Tightly Managed Optical Grid - An Optiputer Vision” (Keynote), Proc. 18th IEEE Int. Conf. on Advanced Information Networking and Application (AINA’04), Vol. 1, pp. 2–11, Fukuoka, Japan, Mar 2004,.
5. D. Feitelson, “Workload modelling for computer systems performance evaluation”, http://www.cs.huji.ac.il/~feit/wlmod
6. E. Medernach, “Workload analysis of a cluster in a Grid environment”, Proc. 11th Int. Workshop Job Scheduling Strategies for Parallel Processing (JSSPP), Cambridge, MA, USA, Jun 2005.
7. H. Li, M. Muskulus, and L. Wolters, “Modeling Job Arrivals in a Data-Intensive Grid”, Proc. 12th Workshop on Job Scheduling Strategies for Parallel Processing, Saint-Malo, France, Jun 2006.
8. T. Stevens, et al., “Anycast Routing Algorithms for Effective Job Scheduling in Optical Grids”, Proc. European Conference on Optical Communication (ECOC), Cannes, France, Sep 2006.
9. W. Fisher, and K. Meier-Hellstern, “The Markov-modulated Poisson process (MMPP) cookbook”, Performance Evaluation, 18(2):149-171, Sep 1993.
10. http://public.eu-egee.org/
11. G. McLachlan, and T. Krishnan, “The EM Algorithm and Extensions”, Wiley Series in Probability and Statistics, 1997.
12. Real Time Monitor, http://gridportal.hep.ph.ic.ac.uk/rtm
13. S. Asmussen, O. Nerman, and M. Olsson, “Fitting phase-type distributions via the EM algorithm”, Scandinavian Journal of Statistics, 23(4):419-441, 1996.
14. W. Roberts, et al., “On Ryden’s EM Algorithm for Estimating MMPPs”, IEEE Signal Processing Letters, 13(6):373-376, Jun 2006.