Fluid computation of the performance–energy trade-off in large scale Markov models

Anton Stefanek Richard A. Hayden Jeremy T. Bradley
Department of Computing, Imperial College London, SW7 2BZ

ABSTRACT

Recent fluid analysis techniques allow fast and efficient calculation of complex reward metrics and passage time probabilities in systems with very large state space. We demonstrate how to incorporate these to look at the trade-off between service level agreement (SLA) satisfaction and complex reward optimisation. We show how the fluid analysis naturally leads to a constrained global optimisation problem with embedded differential equations. We illustrate this problem on an abstract model of a virtualised execution environment that accurately captures resource allocations.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Reliability, availability and serviceability; G.3 [**Markov processes**]:

General Terms

Performance

1. INTRODUCTION

One of the critical trade-offs providers of large scale computing clusters have to deal with is that of running costs such as energy consumption versus the availability and response time of the system as viewed by its users. In case of clusters consisting of a large number of heterogeneous nodes interacting in complex ways with the user submitted jobs, it is hard to predict the effects of different system configurations on the operational costs and the performance offered to the users. Building on existing results in fluid analysis, we present a framework where it is possible to efficiently evaluate both the accumulated cost and response time quantiles. We illustrate the techniques on a model of a cluster with multiple job and server classes that can accurately capture resource allocation and service forwarding. The fluid framework leads to a constrained global optimisation problem with embedded differential equations. We demonstrate how solutions to this problem can address the trade-off in the sample model, by giving decisions about the cluster configuration and scheduling policies that minimise energy consumption while maintaining multiple service level agreements.

2. VIRTUALISED EXECUTION MODEL

Consider an abstract model of a virtualised execution cluster, consisting of a large number of nodes, capable of processing a large number of incoming job requests from its users. The nodes can be of two different classes. The "fast" nodes have high performance and are capable of executing multiple jobs at a time, but have a high energy consumption. The "slow" nodes provide more moderate service at a lower cost. Incoming jobs can originate from users of two priorities, "low" (L) and "high" (H). The provider of such cluster can have different service level agreements (SLA) with users of each priority. We consider SLAs that specify the minimum probability with which each job has to be executed within a given time requirement. For example, it can be agreed that H priority jobs are finished within 6.5 seconds at least 90% of the time and, less strictly, L priority jobs within 8 seconds at least 80% of the time. The scheduler probabilistically decides where to allocate incoming jobs, depending on the user priority and node class and the number of free slots on the node. Figure 1 shows an overview of the system. For a given job workload, the cluster provider has to use a sufficient number of nodes and a suitable scheduling policy so that both SLAs are satisfied. Out of these configurations, the one with the minimum total energy consumption should be chosen.

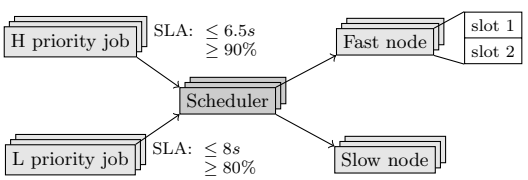

Figure 1: Overview of the cluster model.

3. FLUID ANALYSIS

Clusters like in the example above often consist of more than thousands of nodes. Modelling such systems results in a state space too large for most of traditional analysis techniques. The fluid analysis [3] and mean-field [1] techniques are able to deal with models with such large numbers of identically behaved components. In that case, the state of the system can be described by an integer valued vector keeping track of *populations* of components in each of their state. If the dynamic behaviour of the system can be assumed to be a Markov process, the fluid analysis and mean-field techniques, under certain conditions, derive a set of ordinary differential equations (ODEs) that approximate the time evolution of the populations.

The technique of Hayden *et al.* [3] deals with continuous

time Markov chains coming from models described in the PEPA stochastic process algebra. The numerical solution to the derived systems of ODEs can be used to calculate passage time probabilities, such as the probability of a job finishing within a given time [4]. We have shown how the same system of ODEs can be additionally extended [6] to provide ODEs describing *accumulated rewards* based on the component populations. We can construct linear and non-linear combinations of these rewards to form a good approximation of common cost functions. For example, the total cluster energy consumption can be described as a linear combination of accumulated populations of node states, where the coefficients represent energy profiles of the node in each possible state. Figure 2 shows two examples of a passage time probability and energy consumption derivation from the same sets of ODEs.

The trade-off from the end of the previous section can thus be formulated as the following global optimisation problem with embedded differential equations:

$$\min_{\mathbf{p}} R(\mathbf{x}(t_f, \mathbf{p}), \mathbf{p}) \quad \text{subject to } g_i(\mathbf{x}(t_i, \mathbf{p}), \mathbf{p}) \geq c_i, t_i \in T_c$$
$$\text{where } \dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}) \qquad \mathbf{x}(t_0, \mathbf{p}) = \mathbf{x}_0(\mathbf{p})$$

The vector $\mathbf{x}$ is the population vector with initial values $\mathbf{x}_0$ and $\mathbf{p}$ are parameters such as the number of slow and fast nodes and probabilistic rates determining the scheduling policy. The objective function R is the reward function, for example representing the total energy consumption until time t_f that needs to be minimised. The constraints g_i are functions deriving passage time probabilities from the ODE solution [4] evaluated at the time limits t_i and c_i are the respective SLA thresholds. There is some existing work on solving such optimisation problems [5]. However, in case of models of computer systems the definition of $\mathbf{f}$ can contain the non-smooth minimum function, generated by the contention for resources. So far, we have an efficient implementation that allows naive parameter sweeping and also application of generic approximate global optimisation algorithms, such as MATLAB *Global search*.

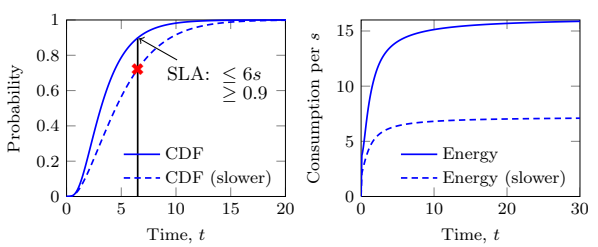

Figure 2: Example of passage time CDF and energy plot. The dashed lines show values from a system with fewer nodes in the cluster.

3.1 Transactions

The original fluid techniques [3] do not accurately capture dependence between different cooperating components. For example there is no easy way of allocating a specific job priority onto a dedicated node class in the PEPA process algebra. If modelled naively, PEPA would only provide the average behaviour of all the nodes serving all the jobs. This has been tackled to some extent by showing how PEPA models can emulate *Layered Queueing Networks* [**?**]; however, this approach does not allow for the possibility of service forwarding which is essential for modelling the job–scheduler–node relationship, where incoming jobs have to be forwarded by the scheduler onto their designated node.

The forwarding and resource allocation cause the system components to enter into local "transactions" [2]. These force components to bind to a cooperating partner for the duration of a multi-stage execution and not, as happens now, allow for arbitrary switching of the partner half way through the execution. The fluid analysis can deal with these by adding new virtual populations of ongoing transactions to the state vector. We can shown how the passage time and reward techniques extend to this class of Markov population models.

4. NUMERICAL EXAMPLES

The virtualised execution cluster model can be described by a Markov population model with transactions. Going through all the possible interactions, the resulting Markov chain has states with 275 components and 734 transition classes. The fluid reward analysis derives a system of 495 ODEs which can be used to formulate a global optimisation problem with 2 SLA constraints. For example, we can fix all the system parameters except for the number of used nodes, with maximum of 60 nodes of each class. If we fix a trivial scheduling where L jobs go only on slow nodes and H jobs go on fast nodes, we get no solution to the problem as no possible configuration can satisfy the demand of L jobs. Another trivial scheduling policy is when the scheduler does not distinguish between priorities and classes and allocates all jobs on all nodes uniformly. The plot on the left of Figure 3 shows the effect of the number of nodes on the SLA satisfaction and energy consumption. Only configurations satisfying both constraint inequalities are shown (the intersection between the two shaded surfaces).

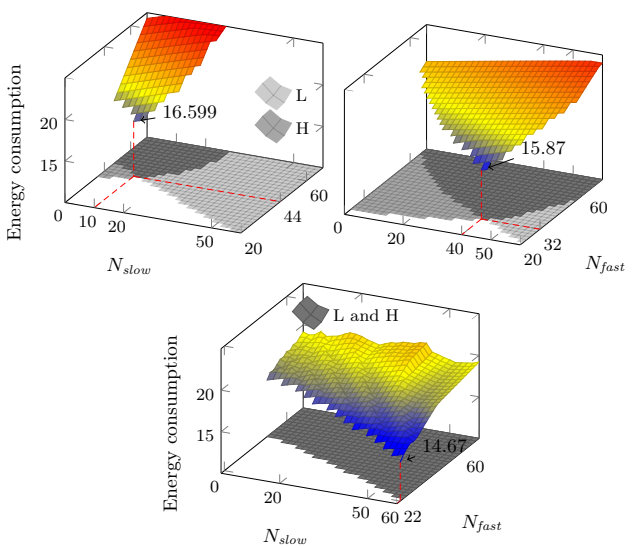

Figure 3: Varying node counts for two scheduling policies and when implicitly optimising scheduling parameters.

A more realistic policy is when a proportion of L jobs is allowed on the fast nodes. The plot on the right of Fig-

ure 3 shows that this leads to an optimal configuration with a lower energy consumption than in case of the uniform scheduling. Finally, the bottom plot in Figure 3 shows for each system configuration the minimum energy consumption obtained by dynamically optimising the scheduling policy. This leads to even lower energy consumption than in case of the static scheduling policies.

5. CONCLUSION

We have demonstrated how a combination of existing fluid techniques can be used to formulate a global optimisation problem. Solution to this problem can address an important trade-off in the design of virtualised execution clusters. By considering transactions of cooperating components, we have shown how such systems can be modelled so that the fluid techniques still apply. We plan to develop more sophisticated optimisation algorithms directly using the ODE structure and extend the techniques to allow a wider range of synchronisation regimes.

6. REFERENCES

[1] N. Gast and G. Bruno. *A mean field model of work stealing in large-scale systems*, volume 38. ACM Press, New York, New York, USA, June 2010.

[2] R. A. Hayden and J. T. Bradley. Shared Transaction Markov Chains for fluid analysis of massively parallel systems. In *2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems*, pages 1–12. IEEE, Sept. 2009.

[3] R. A. Hayden and J. T. Bradley. A fluid analysis framework for a Markovian process algebra. *Theoretical Computer Science*, 411(22-24):2260–2297, May 2010.

[4] R. A. Hayden, A. Stefanek, and J. T. Bradley. Fluid passage-time calculation in large Markov models. *Theoretical Computer Science*, to appear, 2011.

[5] A. B. Singer and P. I. Barton. Global Optimization with Nonlinear Ordinary Differential Equations. *Journal of Global Optimization*, 34(2):159–190, Feb. 2006.

[6] A. Stefanek, R. A. Hayden, and J. T. Bradley. Fluid analysis of energy consumption using rewards in massively parallel markov models. In *Proceeding of the second joint WOSP/SIPEW international conference on Performance engineering*, ICPE '11, pages 121–132, New York, NY, USA, 2011. ACM.