A physics-style approach to scalability of distributed systems

Erik Aurell[1,2], Sameh El-Ansary[1]

[1] Distributed Systems Laboratory
SICS Swedish Institute of Computer Science
P. O. Box 1263 SE-16429 Kista, Sweden

[2]Department of Physics, KTH-Royal Institute of Technology
SE-106 91 Stockholm, Sweden

{eaurell,sameh}@sics.se *

SICS Technical Report T2004:01
ISSN 1100-3154
ISRN:SICS-T–2004/01-SE

Abstract

Is it possible to treat large scale distributed systems as physical systems? The importance of that question stems from the fact that the behavior of many P2P systems is very complex to analyze analytically, and simulation of scales of interest can be prohibitive. In Physics, however, one is accustomed to reasoning about large systems. The limit of very large systems may actually simplify the analysis. As a first example, we here analyze the effect of the density of populated nodes in an identifier space in a P2P system. We show that while the average path length is approximately given by a function of the number of populated nodes, there is a systematic effect which depends on the density. In other words, the dependence is both on the number of address nodes and the number of populated nodes, but only through their ratio. Interestingly, this effect is negative for finite densities, showing that an amount of randomness somewhat shortens average path length.

*This work is funded by the Swedish funding agency VINNOVA, PPC project, and by the European IST-FET PEPITO project.

1 Introduction

In this paper, we propose a method for analyzing properties of large-scale distributed systems based on analogies with thermodynamics and statistical mechanics. That is, given a distributed system of size $\mathcal{N}$ exhibiting a property $\mathcal{P}$, we would like to know the behavior of $\mathcal{P}$ at system sizes $\mathcal{N}$ much greater than where direct simulation is feasible. Particularly we would like to know on the one hand if the description can in any way become simpler for large enough $\mathcal{N}$, or if there is a way to determine what is a "sufficiently large $\mathcal{N}$", so that no simulation of an even larger system is necessary, or is likely to reveal any new information.

Physics was the first science to encounter problems of this sort. The number $\mathcal{N}$ of molecules in a macroscopic body, say a liter of water, is about 10^{27}. On the microscopic level, all substances are made of atoms and molecules of the same basic type – different number of electrons, protons and neutrons – yet large lumps of the same kind of atoms or molecules make up substances with distinct qualities which we can percieve. Those can be density, pressure (of a gas at given density and temperature), viscocity (of a liquid), conductivity (of a metal), hardness (of a solid), how sound and light do or do not propagate, if the material is magnetic, and so on. Materials are important, indeed whole ages of human history have been named after the dominant material at the time [1].

The first level of analysis in a physical system of many components, is to try to separate *intensive* and *extensive* variables. Extensive variables are those that eventually become proportional to the size of the system, such as total energy. Intensive variables, such as density, temperature and pressure, on the other hand becomes independent of system size. A description in terms of intensive variables only is a great step forward, as it holds regardless of the size of the system, if sufficiently large.

Further steps in a physics-style analysis may include identifying phases, in each of which all intensive variables vary smoothly, and where the characteristics of the system remain the same. This methodology was carried over to satisfiability theory more than ten years ago. KSAT is the problem to determine if a conjunction of M clauses, each one a disjunction of K literals out of N variables can be satisfied. Both M and N are extensive variables, while $\alpha = M/N$, the average number of clauses per variable, is an intensive variable. For large N, instances of KSAT fall into either the SAT or the UNSAT phase depending on whether α is larger or smaller than a threshold $\alpha_c(K)$ [3, 2]. The order of the phase transition, a statistical mechanics concept roughly describing how abrupt the transition is, has been shown to be closely related to the average computational complexity of large instances of KSAT with given values of K and α [5, 6]. Recent advances include the introduction of techniques borrowed from the physics of disordered systems, leading to an important new class of algorithms, currently by far the best of large and hard SAT problems [4]. Without question, statistical mechanics have been proven to be very useful on very challenging problems in theoretical computer science, and it can be hoped that this will also be the case in the analysis and design of distributed systems.

2 The physics-style analysis of structured overlays

We start our investigation by considering the area of Peer-To-Peer systems as an example of large-scale distributed systems. The investigation is done using the Chord system [8, 7] and familiarity of the reader to Chord concepts and terminology is assumed.

The work reported in this paper can then be summarized by the three following methodological steps:

Step 1: Determination of intensive variables. Let N be the size of the identifier space and P be the population, i.e. the number of nodes that are uniformly distributed in the identifier space. We define the density (ρ) to be the ratio $\frac{P}{N}$ with a maximum value of 1 for a fully populated system. We will here focus on the investigation of ρ as an intensive variable.

Step 2: Looking for Characteristic Behavior. A key quantity of interest in a P2P system built of DHTs is the average path length. Here we report the dependency of the number of populated nodes, and the density in a series of simulations of Chord. We will show that while the main behaviour is $0.5 \log_2 P$, where P is the number of populated nodes, there is also a small residual term that depends on the density only.

Step 3: Ideas this can give to P2P systems. It is a curious fact that the residual term alluded to above is negative. We call this curious, because if the P populated nodes are regularly spaced in the circular geometry of the address space of Chord, the average path length is exactly $0.5 \log_2 P$, in other words larger. Hence, we have as a result that randomization improves the performance of P2P system built on DHT, even in static situation, with no peers leaving or joining the

```
for $N \in \{2^7, 2^8, .., 2^{14}\}$ do
  for $P \in \{0.1 \times N, 0.2 \times N, .., 1.0 \times N\}$ do
    1. Generate CHORD($P$,$N$)
    2. Inject uniformly distributed $P^2$ lookups
    3. Record the average lookup length over
       the $P^2$ lookups, denoted $< L(P, N) >$ or
       equivalently $< L(\rho, N) >$
  end
end
```

Figure 1: The procedure for investigating the density ($\rho = \frac{P}{N}$) as an intensive variable

system. We believe this may be of some conceptual importance, even if the effect is small.

Further ideas are taken up in the Discussion and Results section below.

3 The simulation set-up

Let CHORD(P,N) be an optimal Chord graph, where all the fingers of all nodes are correctly assigned, our simulations follow the procedure illustrated in figure 1.

This procedure is repeated 10 times, with different random seeds, and the results are averaged.

The simulations were implemented using the Mozart distributed programming platform [9]. The total number of experiments performed was 800 that were scheduled on a cluster of 16 machines at SICS.

4 Results

Given the set of experiments performed as explained in figure 1, we report the results in a number of different ways.

First, as shown in figure 2, for a given set of P nodes, by placing them in identifier spaces of different sizes, the path length is affected. In fact, the path length decreases as the identifier space increases. The graph is based on a subset of the data points where Ps are equal. This graph is mainly to show that the lookup length is not a function of the population alone.

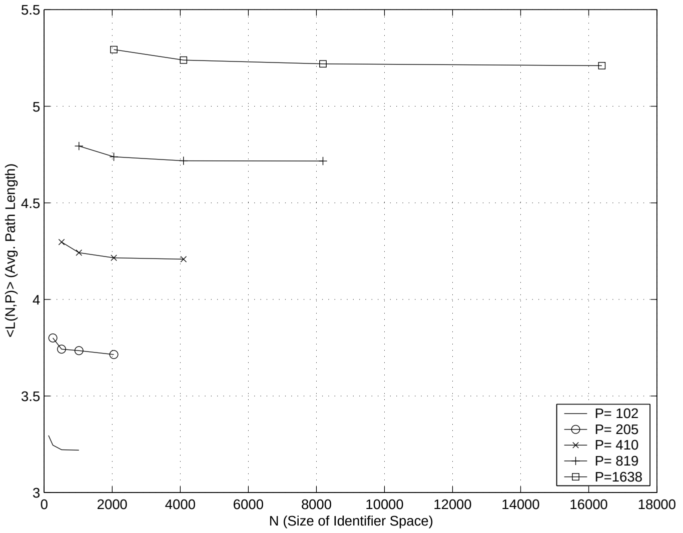

Figure 2: The effect of different identifier space sizes on systems of the same population

Second, in figure 3 we show the behavior of the path length as a function of the density and the size of the identifier space. The curves are, to first approximation, vertically shifter by the same distance, while the values of N used are exponentially spaced. This means that the dependence on N alone (constant P) is logarithmic. Indeed, it was noted in the Chord papers that the average path length is $0.5 \times \log P$. However, we can see an additional observation by looking at the data collapse obtained in figure 4 by subtracting $< L(1, N) >$ from every respective curve $< L(\rho, N) >$ compared to $0.5 \log_2 \rho$. From the data collapse, we can clearly see that $< L(\rho, N) >= 0.5 \log_2 \rho + f(\rho)$ where the function f is a decreasing function. That is for any given number of nodes, the lookup length increases when they are placed in a smaller identifier space. Another way of observing this phenomena is shown in figure 5 where from each data point, the respective $0.5 \log_2 P$ is subtracted.

5 Discussion

A main direct result of this paper is that if one compares two Chord systems, both with P nodes, in address spaces of size N_1 and N_2, $N_1 < N_2$, the average path length in the (P, N_1) system is larger. This is somewhat counter-intuitive, since it suggests that having a larger space to look in speeds up the search (on average). Imagine that in both cases the address space is in fact

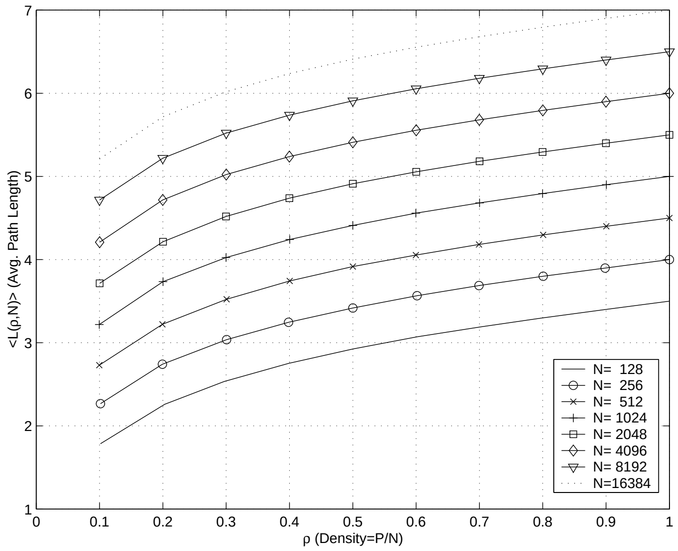

Figure 3: The average lookup length as a function of ρ and N

of the same size N_2, but in the first case only the N_1 regularly spaced nodes can be populated, with gaps of size N_2/N_1 between them.

The explanation is that the Chord routing table has $\log_2 N_2$ elements. If all N_2 addresses can be populated, all entries of the routing table can be used to list a hop that will bring you closer to your destination. In the situation with only N_1 addresses, there is rigidity in the placement of populated nodes (in the address space of size N_2), and all elements in the routing table are not used. Indeed, if written in the N_1 address space, the routing table has but $\log_2 N_1$ entries. Hence, the effect of having more keys to look for locally is slightly stronger than having to search in a larger spaces. In different language, given structured overlay network, it is appearently an advantage to sample it randomly, and not be confined to a subset of keys and possible nodes.

The overall motivation for this work is that physics-style analysis may prove useful in designing and analysing large P2P systems, and we end with a short summary of what we want to accomplish in this direction, and why.

First, it is a well-known fact in the community that simulations of different P2P systems of varying sizes are plentiful, but systematic methods to compare them are more rare. Statistical mechanics is the physical theory of what macroscopic properties that emerge for which microscopic descriptions, and how a large system approaches the infinitely large, or thermodynamic, limit. The ultimate goal when analysing a P2P system is indeed to to find out

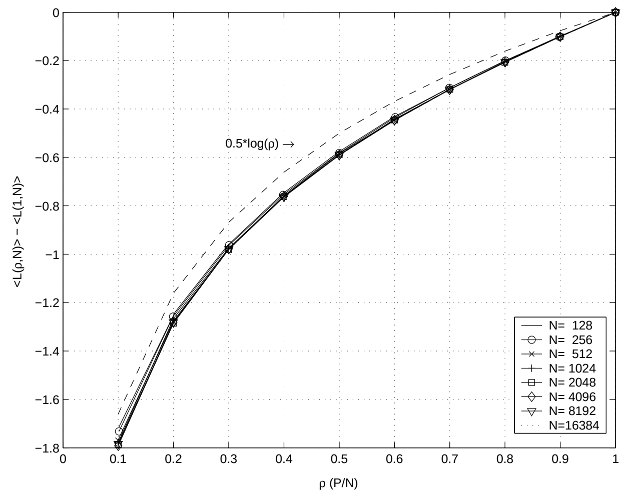

Figure 4: Data collapse of the average lookup length as a function of ρ and N compared to $0.5 \log_2 \rho$

which desirable (or desirable) properties hold for a large system as a whole, when you have but specified the individual component. This is in fact a new statistical mechanics to be discovered. It is also of some practical importance to be able to say that a simulation is large enough, and that nothing essential can be gained by simulating an even larger system. Specific techniques for such tasks are e.g. finite-size scaling.

Second, dynamic properties should also be described by intensive variables. A possibility would be e.g. the average number of join or leave operations per populated node and time between running a stabilization algorithm. One example which has been studied by one of the authors is the ratio of join or leave operations to the number of look-ups generated in DKS [10], a system where updating outdated routing information is performed on the fly. Preliminary results on this system indicate the existence of at least two phases, one "good" (with path length proportional to $\log_2 P$), and one bad (with a very large path length, possibly proportional to P).

Third, it may be a general feature of many P2P systems, that they should preferentiably be operated in a "good" phase, but as close as possible to phase boundaries. This is because goodness usually costs, e.g. in stabilization. Techniques to estimate where phase boundaries lie, and to monitor if they are close in a dynamically changing environment, may hence give new ideas to control of P2P systems. Examples in this direction are spontaneous fluctuations, which grow strongly in size close to at least some types of phase

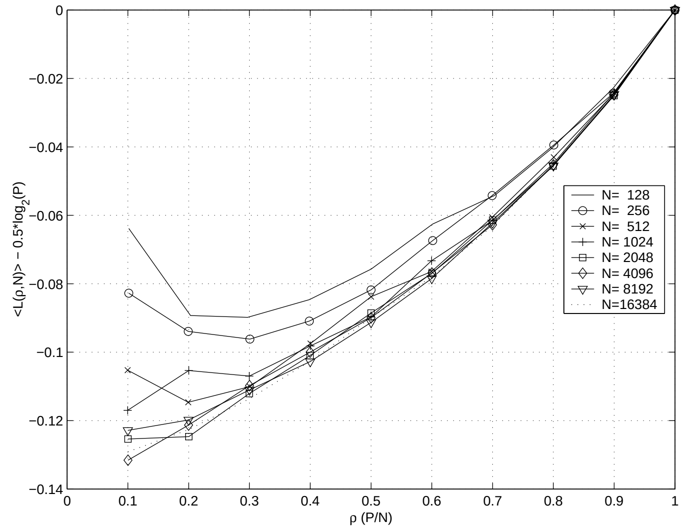

Figure 5: Data collapse by subtracting the respective $0.5 \log_2 P$ from all data points.

transitions.

6 Conclusion

The main contribution of this position paper is the introduction of a methodology for analysing large-scale distributed systems using physical systems analysis techniques. The main goal of such an approach is to illustrate how the behaviors of large systems could be understood while eliminating the need to simulate large instances.

An example of the methodology is provided as a study of a property of the Chord system, namely the density of the population of nodes in an identifier space.

Acknowledgements

We thank Scott Kirkpatrick for discussions on the interface between P2P and statistical physics. We thank Seif Haridi, Per Brand, Luc Onana, Ali Ghodsi and the other members of the DSL laboratory at SICS for many discussions.

References

[1] P. M. Chaikin & T. C. Lubensky, *Principles of condensed matter physics*, Cambridge University Press, 1995.

[2] S. Kirkpatrick & B. Selman, “Critical Behaviour in the Satifiability of Random Boolean Expressions”, *Science* **264** (1994) 1297-1301.

[3] D. Mitchell, B. Selman & H. Levesque, “Hard and easy distributions of sat problems”, in: *Proc. of Am. Assoc. for Artif. Intell.* AAAI-92, 1992, pp. 456–465.

[4] M. Mézard, G. Parisi & R. Zecchina, “Analytic and Algorithmic Solutions of Random Satisfiability Problems”, *Science*, **297**, 812-815 (2002)

[5] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman & L. Troyansky, “Computational complexity from ’characteristic’ phase transitions”, *Nature* **400** (1999) 133–137.

[6] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman & L. Troyansky, “2+p-sat: Relation of typical-case complexity to the nature of the phase transition”, *Random Structures and Algorithms* **3** (1999) 414.

[7] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan, *Chord: A scalable peer-to-peer lookup service for internet applications*, Tech. Report TR-819, MIT, January 2002.

[8] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *ACM SIGCOMM 2001*, pages 149–160, San Deigo, CA, August 2001.

[9] Mozart Consortium. http://www.mozart-oz.org

[10] Luc Onana Alima, Sameh El-Ansary, Per Brand, and Seif Haridi, *Dks(n; k; f): A family of low communication, scalable and fault-tolerant infrastructures for p2p applications*, To appear in the 3rd International workshop on Global and Peer-To-Peer Computing on large scale distributed systems (Tokyo, Japan), May 2003.