Guia do Doutorando

SDDI 2006
Fevereiro de 2006

Doutorando:
José Carlos Rufino Amaro (rufino@ipb.pt)

Orientador:
António Manuel da Silva Pina (pina@di.uminho.pt)

A Identificação

A.1 Doutorando

José Carlos Rufino Amaro (rufino@ipb.pt)
Instituto Politécnico de Bragança, Departamento de Informática e Comunicações

A.2 Título da Tese

Domus:
Tabelas de Hash Distribuídas em Clusters de Nós de Computação Heterogéneos

A.3 Orientador

António Manuel da Silva Pina (pina@di.uminho.pt)
Universidade do Minho, Departamento de Informática

A.4 Data de Início

Data Oficial: 31 de Outubro de 2000; Data Efectiva: 1 de Setembro de 2001

A.5 Data de Término

Data Oficial: 30 de Setembro de 2005; Data Efectiva: 30 de Setembro de 2006

B Resumo

B.1 Área de Investigação e Desenvolvimento

Computação Paralela e Distribuída
Tabelas de Hash Distribuídas em Clusters de Nós de Computação Heterogéneos

B.2 Resumo

[Resumo provisório, baseado nas contribuições principais]

A investigação prosseguida na tese aborda, em termos gerais, a temática das Estruturas de Dados Distribuídas, orientadas a ambientes do tipo Clusters de nós de computação heterogéneos. O trabalho culmina com a definição e desenvolvimento de

uma arquitectura para Tabelas de Hash Distribuídas (DHTs), baseada em propostas de modelos originais, e outros modelos adaptados da literatura, que dá uma resposta adequada a problemas concretos relacionados com a *distribuição*, a *localização*, o *armazenamento* e o *balanceamento dinâmico* de múltiplas DHTs. O resultado final pode ser avaliado em termos das contribuições individuais resumidas a seguir.

A primeira contribuição materializa-se sobre a forma de um modelo para a distribuição pesada de uma DHT, num Cluster de nós computacionais, assente na definição de duas abstracções principais para a subdivisão da DHT: *partições* (de grão fino) e *nós virtuais* (de grão médio). O modelo desenvolvido é compatível com a) a existência de um número variável de nós, b) o dinamismo em termos de grau de participação de cada nó numa DHT, c) a parametrização da aproximação entre a quota ideal e a quota real de participação de cada nó numa DHT e d) a escalabilidade das operações que d1) alteram a composição dos nós envolvidos numa DHT e d2) o respectivo grau de participação, sendo o número global de nós envolvidos nessas operações (bem como de mensagens trocadas) de ordem $O(1)$.

Outra contribuição relevante são os algoritmos de *encaminhamento agregado* desenvolvidos para a localização distribuída de *partições* de DHTs, aplicáveis a topologias *Chord* e *de Bruijn*, baseados na disponibilidade de múltiplas tabelas de encaminhamento em cada nó; os algoritmos permitem reduzir consideravelmente o número médio de saltos de encaminhamento, quando comparados com os algoritmos convencionais, estes baseados na análise de uma só tabela por cada decisão de encaminhamento. Para cada uma das topologias referidas, foram estudadas várias soluções, cada uma deles representando um compromisso específico entre a carga computacional associada a uma decisão de encaminhamento e o número médio de saltos de encaminhamento, por cada localização distribuída de uma partição.

A partir das contribuições anteriores, foi investigada a hipótese de melhorar a eficiência de utilização dos recursos do Cluster, através da atribuição das funções de *endereçamento* e de *armazenamento* de uma mesma partição, de uma DHT, a nós diferentes. O resultado foi a definição da "arquitectura Domus", que suporta a realização, gestão e exploração de múltiplas DHTs num ambiente de Cluster baseado em nós heterogéneos. A arquitectura pressupõe a co-existência dos seus serviços com múltiplas aplicações exógenas, que executam concorrentemente no Cluster, impondo requisitos dinâmicos e imprevisíveis nos recursos dos nós. Neste quadro, a arquitectura prevê mecanismos de ajuste/balanceamento dinâmico de carga que, por um lado, a) procuram rentabilizar, em cada instante, os recursos do Cluster e, por outro lado, b) procuram assegurar níveis esperados de serviço para cada DHT. Cumulativamente, a possibilidade de estabelecer, para cada DHT, um conjunto inicial de atributos, proporciona a sua adequação a requisitos aplicacionais específicos. Outras facilidades, com destaque para a possibilidade de congelamento e reactivação das DHTs, permitem que a execução de certas aplicações paralelas/distribuídas seja intermitente, ou que certos recursos, como RAM, possam ser comutados entre DHTs.

B.3 Objectivos Estratégicos

É objectivo principal desta tese contribuir para o enriquecimento das abordagens às Estruturas de Dados Distribuídas (DDSs) orientadas a ambientes paralelos/distribuídos do tipo Cluster, num quadro de exploração dinâmico e partilhado.

Em particular, pretende-se desenvolver modelos (e, se possível, plataformas), para Tabelas de Hash Distribuídas (DHTs), cientes da heterogeneidade do ambiente de exploração envolvente e auto-ajustáveis ao dinamismo das suas condições, com o objectivo último de optimizar os recursos do Cluster, durante a execução das DHTs.

Mais recentemente, o confronto com a generalização da utilização de Clusters com base em ambientes de trabalho em lotes (vulgo ambientes do tipo *batch*), sugeriu a conveniência de compatibilizar com eles as plataformas entretanto por nós já desenvolvidas, de forma a poderem ser usadas como aplicações de nível utilizador, escalonáveis e executáveis como aplicações regulares paralelas/distribuídas que, por seu turno, oferecem serviços suplementares a outras aplicações clientes.

C Contribuições

C.1 Principais contribuições técnico–científicas

1. um modelo para a distribuição pesada de uma DHT através de um conjunto dinâmico e heterogéneo de nós computacionais, baseado em abstracções de grão fino (*partições*) e médio (*nós virtuais*) para a subdivisão de uma DHT;

2. algoritmos de *encaminhamento agregado*, para a localização distribuída de *partições* de DHTs, aplicáveis a grafos *Chord* e *de Bruijn*, que reduzem o número médio de saltos de encaminhamento, relativamente a métodos convencionais;

3. a "arquitectura Domus", cuja mais valia reside na integração das contribuições anteriores e na atribuição independente e dinâmica, das funções de *endereçamento* e de *armazenamento* de uma partição, baseada na identificação dos nós de computação mais adequados à execução daquelas funções;

4. um protótipo que cumpre a parte mais significativa das funções previstas pela arquitectura e realiza a API definida, sendo de execuçao compatível com modelos de exploração dos tipos *partilhado* e *exclusivo* (i.e., baseados em reservas);

C.1.1 Relevância das Contribuições

A primeira contribuição é relevante no panorama dos modelos de hashing dinâmico distribuído, pelo facto de oferecer, como novidade, a possibilidade de parametrizar

a aproximação entre as quotas ideias e reais, de uma DHT, associadas a cada nó, num quadro em que o número de nós e as suas quotas se admitem dinâmicos.

Os resultados da segunda contribuição são aplicáveis em qualquer cenário que recorra à localização distribuída sobre grafos *Chord* e *de Bruijn* e existe disponibilidade de mais informação de encaminhamento que o habitual (e.g., proveniente de uma vizinhança de nós), o que permite a aceleração da localização distribuída.

A relevância das duas contribuições restantes mede-se pela relativa escassez de abordagens (e realizações) que, em simultâneo, ofereçam a) o suporte integrado a múltiplas DHTs, b) com propriedades eventualmente diferentes, c) com possibilidade de congelamento e reactivação e d) sujeitas a mecanismos de balanceamento dinâmico que permitem a sua utilização em ambientes de exploração partilhados.

No contexto do projecto de investigação em que os trabalhos de doutoramento foram inseridos (projecto SIRe[1]), assume particular importância a contribuição final que se traduz numa plataforma de execução apropriada ao suporte dos vários dicionários de dados necessários, para a execução de robôs paralelos/distribuídos.

C.2 Publicações

1. J. Rufino, A. Pina, A. Alves, and J. Exposto. Distributed Paged Hash Tables (Extended Abstract). International Workshop on Performance-oriented Application Development for Distributed Architectures, University of Technology, Munich, April 19-20, 2001

Resumo

The interest on cluster computing has been growing, in part by the availability of powerful microprocessors and high-speed networks as off-the-shelf commodity components, as well as in part by the emerging software components to support high performance applications. However, in order to develop specific applications, libraries to take full advantage from cluster hardware (avoiding unnecessary overheads, like consistency protocols) are still not available. This paper presents a Distributed Hash?Table System with parallel access, developed using a platform named CoR that intends to exploit cluster tecnologies, including support to high performance communication techniques like Myrinet. The main goal of this system is to take a fair advantage of the primary and secondary memory support from all the cluster hosts. We also describe the implementation of a web robot to evaluate the proposed system.

2. J. Rufino, A. Pina, A. Alves, and J. Exposto. Distributed Paged Hash Tables. In *Proceedings of the 5th International Meeting on High Performance Computing for Computational Science (VECPAR 2002) - Selected Papers and Invited Talks*, pages 679–692, Oporto, Portugal, June 2002. Springer-Verlag

Resumo

In this paper we present the design and implementation of DPH, a storage layer for cluster environments. DPH is a Distributed Data Structure (DDS) based on the distribution of a paged hash table. It combines main memory with file system resources across the cluster in order to implement a distributed *dictionary* that can be used for the storage of very large data sets with key based addressing techniques. The DPH storage layer is supported by a collection of cluster–aware utilities and services. Access to the DPH interface is provided by a user–level API. A preliminary performance evaluation shows promising results.

3. J. Rufino, A. Pina, A. Alves, and J. Exposto. Toward a dynamically balanced cluster oriented DHT. In M. H. Hamza, editor, *Proceedings of the International Conference on Parallel and Distributed Computing and Networks (PDCN'04)*, Innsbruck, Austria, February 2004. Acta Press, ISBN: 0-88986-369-5 (420-1)

Resumo

In this paper, we present a model for a cluster oriented Distributed Hash Table (DHT). It introduces *software nodes*, *virtual nodes* and *partitions* as high level entities that, in conjunction with the definition of a certain number of invariants, provide for the balancement of a DHT across a set of heterogeneous cluster nodes. The model has the following major features: **a)** the share of the hash table handled by each cluster node is a function of its *enrollment level* in the DHT; **b)** the enrollment level of a cluster node in the DHT may change dynamically; **c)** cluster nodes are allowed to dynamically join or leave the DHT. A preliminary evaluation proved that the quality of the balancement of partitions of the hash table across the cluster, measured by the standard deviation with relation to the ideal average, surpass the one achieved by using another well known approach.

4. J. Rufino, A. Alves, A. Pina, and J. Exposto. A cluster oriented model for dynamically balanced DHTs. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS '04)*, Santa Fe, New Mexico, USA, April 2004. IEEE. ISBN 0-7695-2132

Resumo

In this paper, we refine previous work on a model for a Distributed Hash Table (DHT) with support to dynamic balancement across a set of heterogeneous cluster nodes. We present new high-level entities, invariants and algorithms developed to increase the level of parallelism and globally reduce memory utilization.

In opposition to a global distribution mechanism, that relies on complete knowledge about the current distribution of the hash table, we

adopt a local approach, based on the division of the DHT into separated regions, that possess only partial knowledge of the global hash table.

Simulation results confirm the hypothesis that the increasing of parallelism has as counterpart the degradation of the quality of the balancement achieved with the global approach. However, when compared with Consistent Hashing and our global approach, the same results clarify the relative merits of the extension, showing that, when properly parameterized, the model is still competitive, both in terms of the quality of the distribution and scalability.

5. J. Rufino, A. Pina, A. Alves, and J. Exposto. Domus - An Architecture for Cluster-oriented Distributed Hash Tables. In *Proceedings of the 6th International Conferenece on Parallel Processing and Applied Mathematics (PPAM'05)*, Poznan, Poland, September 2005. Springer-Verlag

Resumo

This paper presents a high level description of Domus, an architecture for cluster-oriented Distributed Hash Tables. As a data management layer, Domus supports the concurrent execution of multiple and heterogeneous DHTs, that may be simultaneously accessed by different distributed/parallel client applications. At system level, a load balancement mechanism allows for the (re)distribution of each DHT over cluster nodes, based on the monitoring of their resources, including CPUs, memory, storage and network. Two basic units of balancement are supported: *vnodes*, a coarse-grain unit, and *partitions*, a fine-grain unit. The design also takes advantage of the strict separation of object *lookup* and *storage*, at each cluster node, and for each DHT. Lookup follows a distributed strategy that benefits from the joint analysis of multiple partition-specific routing information, to shorten routing paths. Storage is accomplished through different kinds of data repositories, according to the specificity and requirements of each DHT.

D Enquadramento

A partir de meados da década de 90, assiste-se à emergência de economias de escala ao nível da produção de componentes de hardware que, enquadradas no surgimento e generalização de bibliotecas para passagem de mensagens, tais como o PVM [7] e o MPI [8], viabilizaram a montagem e a exploração de Clusters de estações de trabalho [9], com relações custo/benefício muito atractivas, relegando dessa forma para nichos de utilização muito específicos, os vetustos sistemas (massivamente) paralelos.

Para além de computação distribuída pura para cálculo científico, o suporte a quantidades maciças de dados, através do seu armazenamento distribuído, constitui outra das potencialidades típicas dos Clusters, frequentemente associada à anterior.

Neste contexto, surgiu a necessidade de fazer evoluir as estruturas de dados clássicas (*e.g.*, árvores, tabelas de hash, etc.), concebidas para a realização em sistemas centralizados/paralelos, para responder aos novos desafios colocados pela distribuição (*e.g.*, escalabilidade, tolerância a faltas, etc.), inaugurando-se, desta forma, a investigação em torno da temática das Estruturas de Dados Distribuídas (DDSs).

No domínio das Tabelas de Hash Distribuídas (DHTs), abordagens como a LH* [10], a DDH [11], a EH* [12] ou outras orientadas a suporte de serviços Web [13] representam as contribuições mais significativas de uma primeira geração de DDSs. Nessas DHTs podemos identificar um suporte ausente ou muito limitado à heterogeneidade do ambiente de execução e o recurso, na maioria dos casos, a esquemas de localização centralizados ou pouco escaláveis (mais não seja porque, à data, a dimensão dos maiores Clusters era modesta, quando comparada com os de hoje).

Mais recentemente, emergiu uma segunda geração de DHTs [14, 15, 16, 17], concebida para resolver o problema da localização de objectos em ambientes *Peer-to-Peer* (P2P) [18, 19], ambientes distribuídos à escala planetária, sobre a Internet. Nesses ambientes massivamente distribuídos, os desafios da distribuição são elevados a um novo patamar, num contexto que acaba por divergir do ambiente relativamente controlado e auto-contido do Cluster. A concepção de uma nova abordagem para DHTs em ambiente Cluster não pode, todavia, ignorar as contribuições oriundas dessa investigação. Por exemplo, o número crescente de nós nos Clusters, torna cada vez mais atractivo o recurso a mecanismos como o da localização distribuída.

Neste quadro, uma das motivações fundamentais para a nossa investigação foi a constatação, à data do início, de um suporte ausente/incipiente, das abordagens de então a DHTs, à heterogeneidade do ambiente de execução, resultante da co-existência de nós computacionais com capacidades de base diferentes, ou do dinamismo imposto na utilização dos recursos pelas aplicações que executam no Cluster.

Assim, da admissão de um modelo de exploração partilhado para o Cluster acabou por resultar a motivação para a necessidade de suportar de forma integrada a co-existência de múltiplas DHTs, por forma a rentabilizar a utilização dos recursos do Cluster e viabilizar a execução simultânea de (múltiplas) aplicações clientes que recorrem a Dicionários Distribuídos. O desenvolvimento de abordagens apropriadas a tais requisitos viria a culminar no estudo de mecanismos de balanceamento *resource-aware* [20, 21] e *application-driven* [22], *i.e.*, dependentes da caracterização precisa do ambiente de execução e orientados às especificidades das nossas DHTs.

No plano aplicado, o projecto SIRe [1] fornece o enquadramento perfeito à nossa investigação. O foco do projecto são as questões ligadas à indexação de informação da Web, re-avaliadas num contexto de resolução paralelo/distribuído [23, 24, 25, 26], em que a necessidade de Estruturas de Dados Distribuídas (em particular de Dicionários Distribuídos, como DHTs), emerge de forma natural. O projecto oferece assim um cenário real e útil de teste e validação das soluções por nós preconizadas.

D.1 Bibliografia Principal

D.1.1 Computação em Ambiente Cluster

- Rajkumar Buyya. *High Performance Cluster Computing*. Prentice Hall PTR, 1999
- M. Baker. Cluster Computing White Paper. University of Portsmouth, UK, December 2000
- A. Vrenios. *Linux Cluster Architecture*. Sams, 1st edition, 2002
- J. Sloan. *High Performance Linux Clusters with OSCAR, Rocks, OpenMosix, and MPI*. O'Reilly Media, Inc, 1st edition, 2004

D.1.2 Hashing Dinâmico

- R. Fagin, J. Nievergelt, N. Pippenger, and H.R. Strong. Extendible hashing: a fast access method for dynamic files. *ACM Transactions on Database Systems*, (315-344):315–344, 1979
- W. Litwin. Linear hashing: A new tool for file and table addressing. In *Proceedings of the 6th Conference on Very Large Databases*, pages 212–223, 1980
- R.J. Enbody and H.C. Du. Dynamic Hashing Schemes. *ACM Computing Surveys*, 20(20):85–113, 1988

D.1.3 Estruturas de Dados Distribuídas

- W. Litwin, M.-A. Neimat, and D.A. Schneider. LH*: Linear Hashing for Distributed Files. In *Proceedings of the ACM SIGMOD - International Conference on Management of Data*, pages 327–336, 1993
- R. Devine. Design and implementation of DDH: a distributed dynamic hashing algorithm. In *Proceedings of the 4th Int. Conf. on Foundations of Data Organization and Algorithms*, pages 101–114, 1993
- W. Litwin, M.-A. Neimat, and D.A. Schneider. LH*: A Scalable, Distributed Data Structure. *ACM Transactions on Database Systems*, 21(4):480–525, 1996
- V. Hilford, F.B. Bastani, and B. Cukic. EH* – Extendible Hashing in a Distributed Environment. In *Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference*, 1997

- S.D. Gribble, E.A. Brewer, J.M. Hellerstein, and D. Culler. Scalable, Distributed Data Structures for Internet Service Construction. In *Proceedings of the Fourth Symposium on Operating Systems Design and Implementation*, 2000

- R. Martin, K. Nagaraja, and T. Nguyen. Using Distributed Data Structures for Constructing Cluster-Based Services. In *Proceedings of yhe 1st Workshop on Evaluating and Architecting Systems dependabilitY (EASY)*, 2001

D.1.4 Localização Distribuída

- Z. Liu. Optimal Routing in the De Bruijn Networks. Technical Report RR-1130, INRIA Centre Sophia Antipolis, Valbonne, France, July 1989

- J-C. Bermond, Z. Liu, and M. Syska. Mean Eccentricities of de Bruijn Networks. Technical Report RR-2114, CNRS - Université de Nice-Sophia Antipolis, Valbonne, France, August 1993

- S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A Scalable Content–Addressable Network. In *Proceedings of the ACM SIGCOMM'01*, 2001

- A. Rowstron and P. Druschel. Pastry: Scalable, Distributed Object Location and Routing for Large Peer–to–Peer Systems. In *Proceedings of the 18th IFIP/ACM Int. Conf. on Distributed Systems Platforms*, 2001

- I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balkrishnan. Chord: A Scalable Peer–to–Peer Lookup Service for Internet Applications. In *Proceedings of the ACM SIGCOMM'01*, 2001

- P. Maymounkov and D. Mazieres. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *Proceedings of the 1st Int. Workshop on P2P Systems (IPTPS02)*, 2002

- H. Balakrishnan, M.F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Looking Up Data in P2P Systems. *Communications of the ACM*, 46(2):43–48, 2003

- D. Loguinov, A. Kumar, V. Rai, and S. Ganesh. Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routig Distances and FaultResilience. In *Proccedings of the ACM SIGCOMM'03*. ACM, August 2003

D.1.5 Balanceamento Dinâmico

- R. Vingralek, Y. Breitbart, and G. Weikum. Snowball: Scalable Storage on Networks of Workstations with Balanced Load. *Distributed and Parallel Databases*, 6(2):117–156, 1998

- M. Zaki, S. Parthasarathy, and W. Li. *High Performance Cluster Computing - Volume 1 - Architectures and Systems*, chapter Customized Dynamic Load Balancing, pages 579–604. Prentice Hall PTR, 1999

- J. Nichols and M. Claypool. Performance Evaluation of Load Sharing Policies with PANTS on Beowulf Cluster. In *Proceedings of the ClusterWorld Conference*, 2003

- A. Rao, K. Lakshminarayanan, S. Surana, R. Karp, and I. Stoica. Load Balancing in Structured P2P Systems. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS '03)*, 2003

- K. Aberer, A. Datta, and M. Hauswirth. The Quest for Balancing Peer Load in Structured Peer-to-Peer Systems. Technical report, Distributed Information Systems Laboratory, Ecole Polytechnique Federale de Lausanne, 2003

- J. Faik. *A Model for Resource-Aware Load Balancing on Heterogeneous and Non-Dedicated Clusters*. PhD thesis, Rensselaer Polytechnic Institute, 2005

- J. Teresco, J. Faik, and J. Flaherty. Resource-Aware Scientific Computation on a Heterogeneous Cluster. *Computing in Science and Engineering*, 7(2):40–50, 2005

D.1.6 Outras Referências

- Scalable Information Retrieval environment (SIRe). SAPIENS/POSI project 41739/CHS/2001. http://sire.estig.ipb.pt

- Federico D. Sacerdoti, Mason J. Katz, Matthew L. Massie, and David E. Culler. Wide Area Cluster Monitoring with Ganglia. In *Proceedings of the IEEE Cluster 2003 Conference*, 2003

- A. Martelli. *Python in a Nutshell*. O'Reilly, 2003

- J. Goerzen. *Foundations of Python Network Programming*. APress, 1st edition, August 2004

E Desenvolvimento

E.1 Ano 2001

[Actividades Realizadas]

- desenho, instalação e validação de um Cluster [46];

- familiarização com a plataforma PVM [7];
- reorientação do tema da tese em torno da temática das Estrutura de Dados Distribuídas para ambiente Cluster, e respectivo levantamento bibliográfico [10, 11, 47, 33, 12, 13, 34];
- primeira iteração da "arquitectura Dash" para Tabelas de Hash Paginadas e Distribuídas, incluindo o desenvolvimento de um primeiro protótipo [48];
- apresentação de Resumo Extendido da arquitectura no evento PADDA'01 [2];
- estabilização/depuração do protótipo da arquitectura Dash e realização de testes diversos ao seu desempenho e escalabilidade;
- participação activa na elaboração de uma candidatura ao Programa SAPIENS, consubstanciada no Projecto SIRe (ref. 41739/CHS/2001), recomendado pelo painel de avaliação para financiamento em Dezembro de 2001 [1] ;

E.2 Ano 2002

[Actividades Realizadas]

- consolidação da investigação realizada no ano anterior, no artigo "Distributed Paged Hash Tables", aceite e apresentado na conferência VECPAR 2002 [3];
- estudo comparativo de funções de hash, incluindo a definição de métricas de selecção de funções de hash em funções de certos requisitos aplicacionais [49];
- selecção das estruturas de dados mais adequadas, em cada nó, ao suporte local de Dicionários Distribuídos, considerando como qualidades mais importantes o suporte à ordenação de registos e à concorrência no acesso [50, 51, 52, 53, 54];
- enriquecimento da plataforma Dash com uma implementação de skip-lists [53] em código aberto (SLPC [55]), sobre RAM, concebida para minimizar a taxa de *cache-misses*, tendo-se observado ganhos de desempenho apenas marginais;
- levantamento de sistemas de ficheiros do tipo *journaling* e experiências preliminares com sistemas de ficheiros *raw*, tendo em vista o enriquecimento da plataforma Dash com estruturas de dados locais do tipo *external memory* [56];
- estudo de estratégias de localização distribuída [36, 14, 15, 16, 17, 18, 19], para utilização na arquitectura Dash, resultando a proposta da estratégia BiCH [57];

E.3 Ano 2003

[Actividades Realizadas]

- refinamento da abordagem BiCH, na vertente algorítmica e no tratamento de questões como *caching*, desconexão e replicação, desenvolvimentos que conduziram à elaboração do artigo "BiCH: Binary Consistent Hashing for Cluster oriented Distributed Hash Tables", submetido à conferência HiPC 2003 [58];
- enriquecimento da abordagem BiCH, com mecanismos de encaminhamento acelerado não-convencionais (utilização conjunta de várias tabelas, resumos e caches de encaminhamento), apresentados no artigo "Enhanced Routing in the BiCH cluster oriented DHT", submetido à conferência Cluster 2003 [59];
- término da linha de investigação seguida na abordagem BiCH, no quadro da rejeição dos artigos anteriores, e após identificação, via simulação, de insuficiências imprevistas no modelo teórico; opção por outra linha, baseada no tratamento separado das questões da *distribuição* e da *localização* em DHTs;
- concepção de um novo modelo para a distribuição pesada, de uma DHT, através de um conjunto dinâmico e heterogéneo de nós computacionais, tendo a sua primeira iteração originado o artigo "Toward a dynamically balanced cluster oriented DHT", aceite e apresentado na conferência PDCN 2004 [4];
- refinamento do novo modelo, incrementando o seu nível de paralelismo potencial, investigação que originou o artigo "A cluster oriented model for dinamycally balanced DHTs", submetido e aceite na conferência IPDPS'04 [5];

E.4 Ano 2004

[Actividades Realizadas]

- revisita à temática das estratégias de localização distribuída em DHTs, tendo em vista a identificação de abordagens compatíveis com os modelos de distribuição pesada de DHTs, concebidos anteriormente; desse estudo resultou a selecção e adaptação das topologias *Chord* e *de Bruijn*, aos modelos referidos;
- enriquecimento do *encaminhamento convencional* das topologias *Chord* e *de Bruijn*, com algoritmos de *encaminhamento agregado*, que aceleram a localização distribuída, pela análise conjunta de múltipla informação de encaminhamento, disponível em cada nó, pelo novo modelo de distribuição [60];
- definição de modelo para escolha da topologia e algoritmo de encaminhamento, resultante da caracterização, por simulação, do número médio de saltos de encaminhamento, e esforço computacional associado, por cada algoritmo [61];

- definição inicial de uma arquitectura integrada – a arquitectura Domus – para a realização, exploração e gestão de DHTs, em ambiente Cluster, com suporte à *multiplicidade*, *heterogeneidade* e *balanceamento dinâmico* de DHTs, assente nos modelos de distribuição e localização definidos anteriormente, e descrita no Relatório Técnico "Uma Arquitectura para Múltiplas Tabelas de Hash Distribuídas em Ambiente Cluster", apresentado no evento SDDI 2004 [62].

E.5 Ano 2005

[Actividades Realizadas]

- segunda iteração da arquitectura Domus, por refinamento da anterior, descrita no artigo "Domus - An Architecture for Cluster-oriented Distributed Hash Tables", submetido e apresentado na conferência PPAM 2005 [**?**];
- expansão da arquitectura com o meta-suporte a múltiplas instanciações [63];
- definição rigorosa dos mecanismos de balanceamento da arquitectura [64];
- investigação de plataformas e ferramentas adequadas à implementação de um protótipo da arquitectura, designadamente de bibliotecas de comunicação, armazenamento, linguagens de programação e ambientes de desenvolvimento;
- implementação de protótipo da arquitectura, que implementa a totalidade do interface público, parte dos mecanismos de balanceamento e inclui um serviço de monitorização do Cluster cooperante com o sistema Ganglia [65];

E.6 Ano 2006

[Actividades Previstas]

- validação da arquitectura, através de experimentação com o protótipo;
- redacção do documente da tese de doutoramento;

F Avaliação

No âmbito do trabalho desenvolvido, o modelo de distribuição pesada de DHTs, quando comparado, através da simulação, com o Hashing Consistente, um dos modelos mais proeminentes da literatura [66], revelou vantagens significativas, tal como foi apresentado em [4, 5].

No que diz respeito aos modelos de encaminhamento agregado, e respectivas simulações, apesar das vantagens já identificadas, no contexto deste trabalho, está ainda por realizar a avaliação rigorosa do seu efectivo mérito científico por pares. Encontra-se neste momento em fase de conclusão uma proposta de artigo científico, a submeter para aprovação em conferências internacionais especializadas, que refina os resultados já apresentados em [60, 61].

A dissociação entre funções de endereçamento e armazenamento, em DHTs, não é em si uma novidade absoluta, sendo sugerida pela primeira vez no âmbito de sistemas P2P [67]. Todavia, tanto quanto sabemos, terá sido aplicada por nós ao ambiente Cluster em primeira instância, como instrumento base dos mecanismos dinâmicos de balanceamento previstos pela arquitectura Domus. Convém salientar que o facto de as duas abordagens seguirem diferentes estratégias dificulta a comparação.

Um dos mecanismos de balanceamento da arquitectura Domus tem semelhanças evidentes com um outro mecanismo recente (ver [68]), no que diz respeito à caracterização dos nós computacionais e à utilização de modelos lineares para distribuição de carga. Todavia, aquele mecanismo é orientado preferencialmente a aplicações de cálculo científico, o que dificulta a avaliação dos seus méritos face à nossa abordagem. Uma outra diferença ao nível do balanceamento tem a ver com o facto do algoritmo usado na abordagem citada, entrar em consideração apenas com os recursos de CPU e de Rede, ignorando o espaço de armazenamento (em RAM ou Disco) e o nível de serviço do Disco, recursos tomados em conta, no nosso caso.

No actual momento de trabalho, para além do desenvolvimento de estudos comprovativos da eficácia dos mecanismos de balanceamento dinâmico previstos pela arquitectura, faltam ainda conduzir experimentos que avaliem a usabilidade do protótipo desenvolvido, recorrendo à biblioteca Domus. Neste contexto, assume particular relevância como plataforma de experimentação, a possibilidade de utilização, pelos investigadores associados ao projecto SIRe [1], dos recursos de computação, de armazenamento e de comunicação de elevado desempenho, integrados no Cluster *Search* sediado no departamento de Informática da Universidade do Minho.

G Dificuldades

O plano de doutoramento iniciado, estabelecido no âmbito do programa Prodep, foi confrontado com um certo número de imprevistos, tanto de carácter pessoal, como profissional e de investigação, que afectaram significativamente o desenvolvimento dos trabalhos, tendo como resultado um adiamento considerável face aos prazos e programa pré-estabelecidos. Enumeram-se, de seguida, os factores mais relevantes.

G.1 Factores Científicos e Técnicos

- necessidade de adaptação e redefinição do tema inicial da dissertação;
- existência de um vasto leque de questões científicas e tecnológicas que obrigaram ao estudo e investigação de temas não previstos inicialmente;
- inadequação e degradação, ao longo do tempo de desenvolvimento dos trabalhos, dos recursos de computação necessários; espera-se que o ambiente *search*, recentemente instalado no Departamento de Informática, venha a contribuir para aliviar o problema;

G.2 Outros Factores

- a dispensa de serviço docente, no Instituto Politécnico de Bragança, ao abrigo do programa Prodep, só foi concretizada a partir do último semestre de 2001;
- responsabilidades familiares acrescidas, desde 2003, vieram diminuir de forma significativa o envolvimento no trabalho;
- o regresso à actividade laboral, terminada a bolsa do Prodep, limitou ainda mais a disponibilidade do doutorando, desde o último semestre de 2004;
- razões de ordem geográfica condicionaram uma aproximação mais frequente do doutorando ao orientador e à Universidade;

Referências

[1] Scalable Information Retrieval environment (SIRe). SAPIENS/POSI project 41739/CHS/2001. http://sire.estig.ipb.pt.

[2] J. Rufino, A. Pina, A. Alves, and J. Exposto. Distributed Paged Hash Tables (Extended Abstract). International Workshop on Performance-oriented Application Development for Distributed Architectures, University of Technology, Munich, April 19-20, 2001.

[3] J. Rufino, A. Pina, A. Alves, and J. Exposto. Distributed Paged Hash Tables. In *Proceedings of the 5th International Meeting on High Performance Computing for Computational Science (VECPAR 2002) - Selected Papers and Invited Talks*, pages 679–692, Oporto, Portugal, June 2002. Springer-Verlag.

[4] J. Rufino, A. Pina, A. Alves, and J. Exposto. Toward a dynamically balanced cluster oriented DHT. In M. H. Hamza, editor, *Proceedings of the International Conference on Parallel and Distributed Computing and Networks (PDCN'04)*, Innsbruck, Austria, February 2004. Acta Press, ISBN: 0-88986-369-5 (420-1).

[5] J. Rufino, A. Alves, A. Pina, and J. Exposto. A cluster oriented model for dynamically balanced DHTs. In *Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS '04)*, Santa Fe, New Mexico, USA, April 2004. IEEE. ISBN 0-7695-2132.

[6] J. Rufino, A. Pina, A. Alves, and J. Exposto. Domus - An Architecture for Cluster-oriented Distributed Hash Tables. In *Proceedings of the 6th International Conferenece on Parallel Processing and Applied Mathematics (PPAM'05)*, Poznan, Poland, September 2005. Springer-Verlag.

[7] Al Geist, A. Beguelin, J. Dongarra, W. Jiang, R. Manchek, and V. Sunderam. *PVM: Parallel Virtual Machine. A User's Guide and Tutorial for Networked Parallel Computing.* Scientific and Engineering Computation. MIT Press, 1994.

[8] M. Snir, S. Otto, S. Huss-Lederman, David Walker, and J. Dongarra. *MPI - The Complete Reference.* Scientific and Engineering Computation. MIT Press, 1998.

[9] M. Baker. Cluster Computing White Paper. University of Portsmouth, UK, December 2000.

[10] W. Litwin, M.-A. Neimat, and D.A. Schneider. LH*: Linear Hashing for Distributed Files. In *Proceedings of the ACM SIGMOD - International Conference on Management of Data*, pages 327–336, 1993.

[11] R. Devine. Design and implementation of DDH: a distributed dynamic hashing algorithm. In *Proceedings of the 4th Int. Conf. on Foundations of Data Organization and Algorithms*, pages 101–114, 1993.

[12] V. Hilford, F.B. Bastani, and B. Cukic. EH* – Extendible Hashing in a Distributed Environment. In *Proceedings of the COMPSAC '97 - 21st International Computer Software and Applications Conference*, 1997.

[13] S.D. Gribble, E.A. Brewer, J.M. Hellerstein, and D. Culler. Scalable, Distributed Data Structures for Internet Service Construction. In *Proceedings of the Fourth Symposium on Operating Systems Design and Implementation*, 2000.

[14] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker. A Scalable Content–Addressable Network. In *Proceedings of the ACM SIGCOMM'01*, 2001.

[15] A. Rowstron and P. Druschel. Pastry: Scalable, Distributed Object Location and Routing for Large Peer–to–Peer Systems. In *Proceedings of the 18th IFIP/ACM Int. Conf. on Distributed Systems Platforms*, 2001.

[16] I. Stoica, R. Morris, D. Karger, M.F. Kaashoek, and H. Balkrishnan. Chord: A Scalable Peer–to–Peer Lookup Service for Internet Applications. In *Proceedings of the ACM SIGCOMM'01*, 2001.

[17] P. Maymounkov and D. Mazieres. Kademlia: A Peer-to-Peer Information System Based on the XOR Metric. In *Proceedings of the 1st Int. Workshop on P2P Systems (IPTPS02)*, 2002.

[18] D. S. Milojicic, V. Kalogeraki, R. Lukose, K. Nagaraja, J. Pruyne, B. Richard, S. Rollins, and Z. Xu. Peer-to-Peer Computing. Technical Report HPL-2002-57, HP Labs, 2002.

[19] H. Balakrishnan, M.F. Kaashoek, D. Karger, R. Morris, and I. Stoica. Looking Up Data in P2P Systems. *Communications of the ACM*, 46(2):43–48, 2003.

[20] J. Faik. *A Model for Resource-Aware Load Balancing on Heterogeneous and Non-Dedicated Clusters*. PhD thesis, Rensselaer Polytechnic Institute, 2005.

[21] J. Teresco, J. Faik, and J. Flaherty. Resource-Aware Scientific Computation on a Heterogeneous Cluster. *Computing in Science and Engineering*, 7(2):40–50, 2005.

[22] Rajkumar Buyya, editor. *High Performance Cluster Computing - Volume 1 - Architectures and Systems*. Prentice Hall PTR, 1999.

[23] A. Heydon and M. Najork. Mercator: A Scalable, Extensible Web Crawler. *World Wide Web*, 2(4):219–229, 1999.

[24] M. Najork and A. Heydon. High-Performance Web Crawling. Technical Report 173, COMPAQ Systems Research Center, September 2001.

[25] V. Shkapenyuk and T. Suel. Design and Implementation of a High–Performance Distributed Web Crawler. Technical Report TR-CIS-2001-03, Polytechnic University, Brooklyn, NY, July 2001.

[26] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. UbiCrawler: A Scalable Fully Distributed Web Crawler. In *Proceedings of the AUSWEB 2002*, 2002.

[27] Rajkumar Buyya. *High Performance Cluster Computing*. Prentice Hall PTR, 1999.

[28] A. Vrenios. *Linux Cluster Architecture*. Sams, 1st edition, 2002.

[29] J. Sloan. *High Performance Linux Clusters with OSCAR, Rocks, OpenMosix, and MPI*. O'Reilly Media, Inc, 1st edition, 2004.

[30] R. Fagin, J. Nievergelt, N. Pippenger, and H.R. Strong. Extendible hashing: a fast access method for dynamic files. *ACM Transactions on Database Systems*, (315-344):315–344, 1979.

[31] W. Litwin. Linear hashing: A new tool for file and table addressing. In *Proceedings of the 6th Conference on Very Large Databases*, pages 212–223, 1980.

[32] R.J. Enbody and H.C. Du. Dynamic Hashing Schemes. *ACM Computing Surveys*, 20(20):85–113, 1988.

[33] W. Litwin, M.-A. Neimat, and D.A. Schneider. LH*: A Scalable, Distributed Data Structure. *ACM Transactions on Database Systems*, 21(4):480–525, 1996.

[34] R. Martin, K. Nagaraja, and T. Nguyen. Using Distributed Data Structures for Constructing Cluster-Based Services. In *Proceedings of yhe 1st Workshop on Evaluating and Architecting Systems dependabilitY (EASY)*, 2001.

[35] Z. Liu. Optimal Routing in the De Bruijn Networks. Technical Report RR-1130, INRIA Centre Sophia Antipolis, Valbonne, France, July 1989.

[36] J-C. Bermond, Z. Liu, and M. Syska. Mean Eccentricities of de Bruijn Networks. Technical Report RR-2114, CNRS - Université de Nice-Sophia Antipolis, Valbonne, France, August 1993.

[37] D. Loguinov, A. Kumar, V. Rai, and S. Ganesh. Graph-Theoretic Analysis of Structured Peer-to-Peer Systems: Routig Distances and FaultResilience. In *Proccedings of the ACM SIGCOMM'03*. ACM, August 2003.

[38] R. Vingralek, Y. Breitbart, and G. Weikum. Snowball: Scalable Storage on Networks of Workstations with Balanced Load. *Distributed and Parallel Databases*, 6(2):117–156, 1998.

[39] M. Zaki, S. Parthasarathy, and W. Li. *High Performance Cluster Computing - Volume 1 - Architectures and Systems*, chapter Customized Dynamic Load Balancing, pages 579–604. Prentice Hall PTR, 1999.

[40] J. Nichols and M. Claypool. Performance Evaluation of Load Sharing Policies with PANTS on Beowulf Cluster. In *Proceedings of the ClusterWorld Conference*, 2003.

[41] A. Rao, K. Lakshminarayanan, S. Surana, R. Karp, and I. Stoica. Load Balancing in Structured P2P Systems. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS '03)*, 2003.

[42] K. Aberer, A. Datta, and M. Hauswirth. The Quest for Balancing Peer Load in Structured Peer-to-Peer Systems. Technical report, Distributed Information Systems Laboratory, Ecole Polytechnique Federale de Lausanne, 2003.

[43] Federico D. Sacerdoti, Mason J. Katz, Matthew L. Massie, and David E. Culler. Wide Area Cluster Monitoring with Ganglia. In *Proceedings of the IEEE Cluster 2003 Conference*, 2003.

[44] A. Martelli. *Python in a Nutshell*. O'Reilly, 2003.

[45] J. Goerzen. *Foundations of Python Network Programming*. APress, 1st edition, August 2004.

[46] A. Alves, J. Rufino, and J. Exposto. Instalação e Configuração de um Cluster sob Linux. Technical report, Instituto Politécnico de Bragança, 2001.

[47] R. Vingralek, Y. Breitbart, and G. Weikum. Distributed File Organization with Scalable Cost/Performance. In *Proceedings of the ACM SIGMOD - International Conference on Management of Data*, 1994.

[48] J. Rufino. Tabelas de Hash Paginadas e Distribuídas. Technical report, Instituto Politécnico de Bragança, 2001.

[49] J. Rufino. Selecção de Funções de Hash. Technical report, Dep. de Informática e Comunicações, Inst. Politécnico de Bragança, Portugal, 2002.

[50] A. Tharp. *File Organization and Processing*. John Wiley & Sons, 1988.

[51] D. Lomet. Grow and Post Index Trees: Roles, Techniques and Future Potential. In *Proceedings of the Second International Symposium on Advances in Spatial Databases*, pages 183–206, 1991.

[52] M. Folk. *File Structures*. Addison-Wesley, 2nd edition, 1992.

[53] W. Pugh. Skip Lists: A Probabilistic Alternative to Balanced Trees. In *Proceedings of the 1st Workshop on Algoritihms and Data Structures (WADS)*, volume 382 of *Lecture Notes in Computer Science*, pages 437–449. Springer-Verlag, August 1999.

[54] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to Algorithms*. The MIT Press, 2nd edition, 2001.

[55] J. MacDonald. Cache-Optimized Concurrent Skip List (SLPC). http://freshmeat.net/projects/skiplist/, 2001.

[56] J.S. Vitter. Online Data Structures in External Memory. In *Proceedings of the 26th Annual Intern. Colloquium on Automata, Languages, and Programming*, 1999.

[57] J. Rufino. Hashing Consistente Binário para uma DHR orientada ao Cluster. Technical report, Instituto Poltiécnico de Bragança, 2002.

[58] J. Rufino, A. Pina, A. Alves, and J. Exposto. BiCH: Binary Consistent Hashing for Cluster oriented Distributed Hash Tables. 2003. (submitted to HiPC 2003; available on request).

[59] J. Rufino, A. Pina, A. Alves, and J. Exposto. Enhanced Routing in the BiCH cluster oriented DHT. 2003. (submitted to Cluster 2003; available on request).

[60] J. Rufino, A. Pina, A. Alves, and J. Exposto. Agregated routing for a cluster oriented DHT. Technical report, Dep. of Informatics and Communications, Polytechnic Institute of Bragança, Portugal, 2004.

[61] J. Rufino, A. Pina, A. Alves, and J. Exposto. Space-time tradeoffs in routing overlays for a cluster oriented DHT. Technical report, Dep. of Informatics and Communications, Polytechnic Institute of Bragança, Portugal, 2004.

[62] J. Rufino. Uma Arquitectura para Múltiplas Tabelas de Hash Distribuídas em Ambiente Cluster. Technical report, Instituto Politécnico de Bragança, 2004.

[63] J. Rufino. Meta-Gestão na Arquitectura Domus. Technical report, Instituto Politécnico de Bragança, 2005.

[64] J. Rufino. Gestão Dinâmica de Carga na Arquitectura Domus. Technical report, Instituto Politécnico de Bragança, 2005.

[65] J. Rufino. Memória Descritiva do Protótipo Domus. Technical report, Instituto Politécnico de Bragança, 2005.

[66] D. Karger, E. Lehman, F. Leighton, D. Levine, and R. Panigrahy. Consistent Hashing and random trees: Distributed caching protocols for relieving hot spots on the world wide web. In *Proceedings of the 29th Annual ACM Symposium on Theory of Computing*, 1997.

[67] K. Chen, L. Naamani, and K. Yehia. miChord: Decoupling Object Lookup from Storage in DHT-Based Overlays. 2003.

[68] J. Faik, J. Flaherty, L. Gervasio, J. Teresco, and K. Devine. A Model for Resource-Aware Load Balancing on Heterogeneous and Non-Dedicated Clusters. Technical Report CS-05-01, Williams College Department of Computer Science, 2005.