A Esquina das Listas: Experiências e Perspectivas

Marcos Euzébio
Ricardo Anido *
Instituto de Computação
UNICAMP
Caixa Postal 6176, 13081-970 Campinas, SP
{euzebio,ranido}@dcc.unicamp.br

Sumário

Muitos serviços da Internet são providos por uma rede virtual de servidores e clientes realizada na malha física de comunicação que a Internet fornece. Assim, o desempenho, organização e gerenciamento de um serviço depende do bom acoplamento entre arquitetura da rede (virtual) do serviço com a própria arquitetura (real) da Internet. A Esquina das Listas foi uma experiência de distribuição de listas que chegou a possuir mais de 12.000 usuários espalhados por quarenta países. Para atender ao desafio imposto pela alta demanda alcançada pelo serviço foi necessário procurar soluções inovadoras de gerenciamento e transporte de listas. Neste artigo é apresentado uma descrição das soluções encontradas e ao mesmo tempo uma reflexão sobre suas virtudes e limitações. À luz desta experiência, é discutida a viabilidade de estender estas soluções visando preparar um arcabouço, com cobertura global na Internet, para o transporte de listas.

Abstract

Many services on the Internet are implemented by a virtual network of servers and clients overlayed on the communication matrix provided by the Internet. The performance, organization and management of such services are heavily dependent on the good coupling between the service (virtual) network and the Internet (actual) architecture. The Esquina das Listas experience reached about 12,000 subscribers spread around forty countries. In order to be able to raise to the challenge imposed by such a high demand on the service it was necessary to find new ways to manage and organize the system. This article will describe some of the solutions found while reflecting about its virtues and limitations. On the ligth of the experience gained it is also discussed the feasibility of extending such solutions in order to create a global framework for the transport of lists.

1 Introdução

Pode-se dizer que a Internet, ao se aproximar de seu trigésimo aniversário, interligando quase 30 milhões de computadores e 120 milhões de usuários [Wiz97], é hoje um sucesso inquestionável. Talvez uma das principais razões para este sucesso tenha sido a oferta de um leque variado de aplicações de comunicação, colaboração e compartilhamento de informação. Estas transformaram a Internet de uma simples plataforma de comunicação em um poderoso instrumento para a cooperação

*Trabalho parcialmente financiado com recursos do CNPq e Fapesp.

entre grupos. De acordo com os dados coletados pelo Laboratório Nacional de Pesquisa Aplicada em Redes (NLANR), mais de 80% dos fluxos, pacotes e bytes circulando pela Internet hoje [WC96] devem-se a este tipo de aplicação, principalmente o WWW.

Por outro lado a Internet vem, já há algum tempo, dando mostras de que este crescimento exponencial é insustentável a longo prazo. Os problemas de adaptação, provocados pelo enorme aumento de escala, têm se manifestado de várias formas e em várias dimensões.

Um tipo de alternativa que tem sido utilizada é a solução por hardware: linhas e computadores mais rápidas e poderosos (também apresentando um crescimento exponencial) poderiam ser adquiridos e seriam capazes de contrapor ao aumento da demanda. Esta alternativa, sem dúvida, quando viável técnica e economicamente, tem sido a adotada e muitas vezes tem conseguido manter a situação dentro do limite do tolerável. Outro tipo de alternativa, complementando as soluções por hardware, seria a solução por software. A reestruturação das aplicações existentes na internet e/ou a introdução de novas aplicações na Internet poderiam melhorar a utilização de seus recursos caros e escassos.

Por exemplo, quando a Internet superou a marca dos 200.000 computadores interligados ficou impraticável a manutenção do sistema de resolução de nomes baseado em um arquivo que era replicado em todos os computadores. A solução foi a introdução do DNS [AL97], que permitiu a implementação de um sistema distribuído para o mapeamento entre nomes e endereços IP. Esta solução resolveu não só o problema técnico de replicação e disseminação da base de nomes e endereços IP mas também os problemas acessórios de gerenciamento, consistência e controle desta base.

É nesta linha que se inclui as contribuições apresentadas neste artigo: a partir das experiências obtidas na implementação e instalação da Esquina das Listas, que chegou a ser o maior servidor de listas brasileiro, oferece-se uma proposta de discussão sobre a viabilidade da implantação de um arcabouço global para o transporte de listas.

O artigo se desenvolve da seguinte forma. Nas duas próximas seções serão apresentados alguns conceitos e definições relacionados à arquitetura da Internet e à estrutura das aplicações encontradas na mesma. A seção seguinte estará dedicada a relatar as experiências obtidas com da Esquina das Listas, principalmente no que se relaciona a aspectos de sua organização e implementação. A proposta para a discussão de um arcabouço vem a seguir. Na última seção serão apresentadas as conclusões do artigo e mencionadas algumas questões pendentes para aprofundamento futuro.

2 A Arquitetura da Internet

Em um sistema paralelo ou distribuído a arquitetura da plataforma de comunicação é fundamental para a estruturação otimizada das aplicações a serem implementadas [Lev87]. O bom desempenho da aplicação depende do ajuste adequado entre a arquitetura da plataforma física de comunicação do sistema e a arquitetura da interconexão lógica da aplicação. Busca-se dividir a aplicação entre os vários componentes do sistema de modo a otimizar o consumo dos recursos de computação e comunicação do sistema.

O crescimento explosivo, continuado e sustentado da Internet, nos últimos anos, é prova da qualidade técnica do seu projeto, capaz de acomodar modificações estruturais e inovações tecnológicas bastante substanciais. Entretanto a aceitação da Internet não pode ser explicada apenas em termos de tais vantagens técnicas. A estratégia administrativa descentralizada, adotada na Internet, com certeza merece uma parcela significativa do crédito pela sua expansão.

Por outro lado, a não existência de uma autoridade central que a organize e a coordene acaba

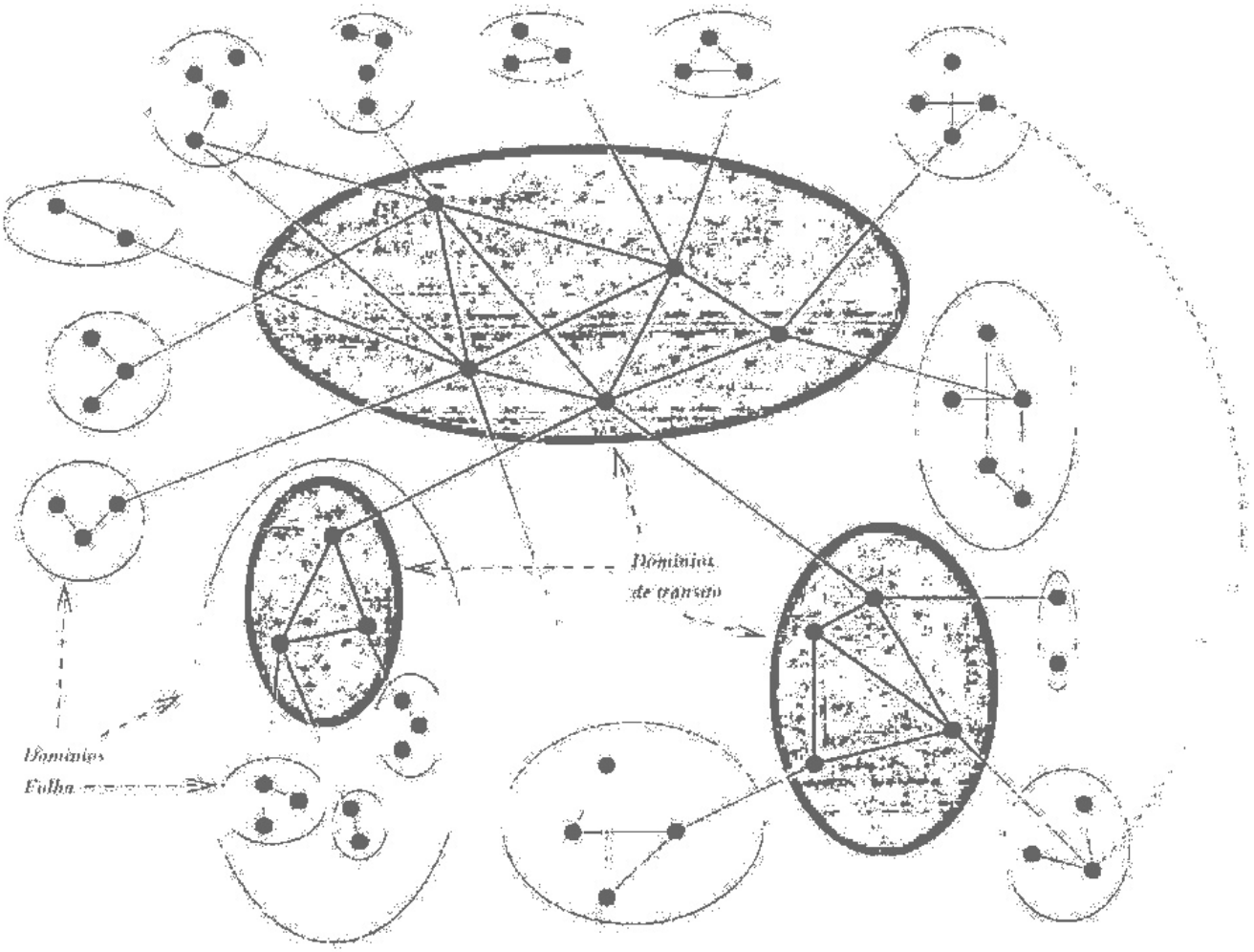

Figura 1: Esboço da arquitetura da Internet

tornando impossível a obtenção de dados precisos sobre a arquitetura da Internet. É bem verdade que sua magnitude, chegando hoje a quase 30 milhões de computadores conectados [Wiz97] e 120 milhões de usuários, e a taxa de crescimento da Internet, dobrando de tamanho a cada ano, já são motivos mais do que suficientes para tornar impraticável qualquer empreitada nesta direção.

No entanto, a maneira descentralizada de administração da Internet oferece algumas pistas para a elaboração de um esboço qualitativo da sua arquitetura. Deste esboço emerge o padrão da arquitetura da Internet [CDZ97]: a Internet pode ser vista como uma quase-hierarquia composta pela interconexão de vários domínios de roteamento. Cada um destes domínios é administrado por uma autoridade autônoma e independente responsável pela implementação de políticas próprias de administração e roteamento.

Os domínios de roteamento podem ser classificados em domínios de trânsito e domínios folhas. Os primeiros permitem o passagem de pacotes originados e destinados a outros domínios. Isto não acontece nos domínios folhas que, porém, podem estar conectados a mais de um domínio de trânsito. Podem existir também conexões dedicadas a carregar o tráfego local entre dois domínios folhas. Um domínio folha, por sua vez, pode ser estruturado em termos de sub-domínios de trânsito e folhas. Estes conceitos estão ilustrados na figura 1.

É bem verdade que, para os usuários da Internet, toda esta irregularidade e complexidade acaba desaparecendo debaixo das diversas camadas de software que cobrem o hardware da Inter-

net. Dentre os vários serviços oferecidos por cada uma das camadas IP e superiores está o de prover uma abstração de rede virtual completa, de modo a permitir a comunicação direta entre as entidades clientes daquela camada. Ou seja, cada uma destas camadas deve tornar transparentes aspectos relacionados ao roteamento da comunicação entre as entidades clientes, liberando-as desta preocupação.

Por exemplo, fazendo um corte horizontal na Internet, acima da camada IP vê-se uma rede virtual completa no nível de computadores hospedeiros *(hosts)*. Um corte acima da camada TCP oferece a visão de uma rede virtual completa no nível de portas de comunicação. Se sobre esta camada coloca-se uma camada SMTP, de servidores de correio eletrônico, tem-se uma rede virtual completa no nível de caixas postais. Mas se a opção for pela camada HTTP, dos servidores WWW, então a rede é completa no nível de documentos hipermídia.

A estruturação vertical da Internet em camadas e protocolos associados, que obriga a passagem de uma mensagem pelo tratamento e processamento de vários estágios de processamentos locais, nos roteadores e computadores envolvidos em uma comunicação, implica em um custo razoável. Van Renesse e colegas, por exemplo, chegaram a detectar a elevação de uma ordem de magnitude na latências de mensagens usadas para o processamento de RPC, no contexto de uma rede local [VRST88]. Ou seja, o *overhead* de processamento acaba dominando o tempo gasto pela mensagem no cabo. Resultados similares também tem sido observados em gerenciadores de listas: gasta-se mais tempo aguardando que uma conexão se complete do que transmitindo a mensagem.

3 Aplicações na Internet: Organização e Arquiteturas Típicas

Como estruturar uma aplicação que será implantada sobre a plataforma de comunicação oferecida pela Internet? O fato desta plataforma possuir uma arquitetura virtual completa dá ao desenvolvedor da aplicação uma ampla liberdade de escolha.

Em geral estas aplicações são desenvolvidas de acordo com o paradigma cliente-servidor, ou seja, aspectos ligados à interface da aplicação são implementados no lado do cliente e aspectos ligados à funcionalidade efetiva da aplicação no lado do servidor. A arquitetura da aplicação neste caso é uma estrela: o servidor ao centro conectado aos clientes nas pontas, conforme mostrado na figura 2. Ao longo do tempo o servidor poderá atender a vários clientes simultaneamente e/ou consecutivamente.

Quando apenas um servidor é insuficiente para atender as necessidades do conjunto de clientes a solução é ter mais servidores. A partir deste ponto várias alternativas são possíveis e dependem do tipo da aplicação. Cada alternativa determina diferentes graus de integração, coordenação e comunicação entre os servidores.

3.1 Sistemas Dispersos e Distribuídos

Se o domínio da aplicação pode ser fragmentado em vários sub-domínios independentes a solução é fazer uma replicação funcional dos servidores. Cada um dos servidores poderá atuar de forma autônoma. Também é certo que nem sempre esta fragmentação pode ser feita de forma transparente para os clientes do serviço e alguma maneira, interna ou externa à aplicação, tem de ser encontrada para colocar o cliente em contato com o servidor encarregado pelo sub-domínio adequado da aplicação.

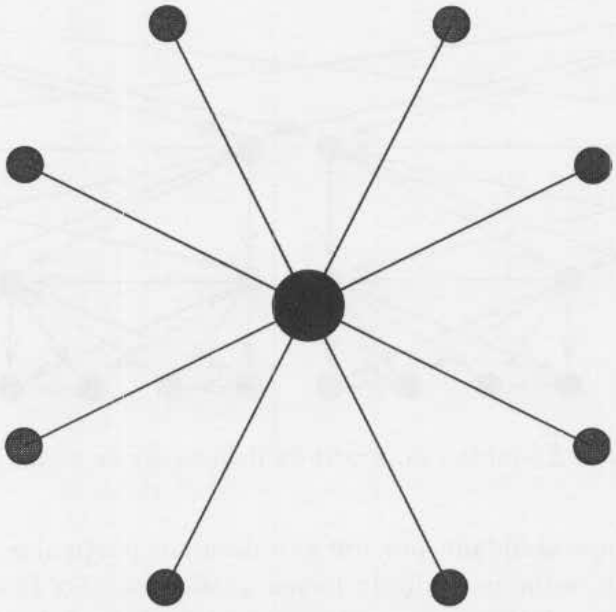

Figura 2: Arquitetura de uma aplicação cliente-servidor simples

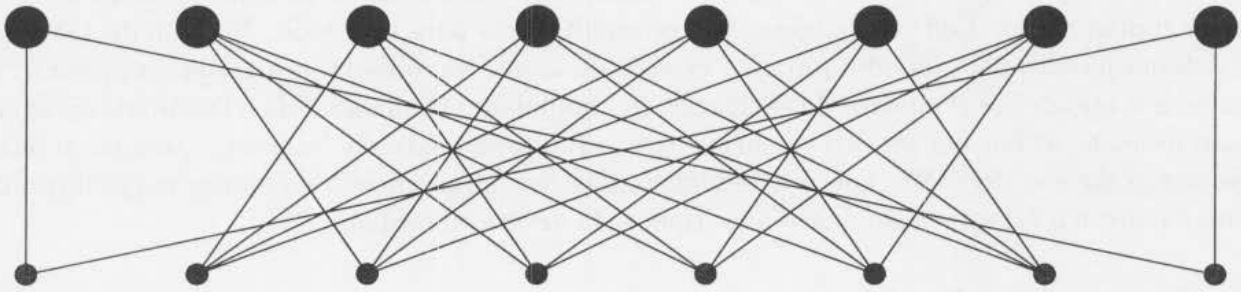

Figura 3: Arquitetura de um sistema disperso

3.1.1 Sistemas Dispersos

Quando não existe uma necessidade maior de integração entre os servidores a aplicação forma um **sistema disperso.** Todo servidor pode seguir a melhor política de administração e controle que convém à autoridade responsável pelo mesmo. Nos sistemas dispersos os conjuntos combinados de clientes e servidores, associados à aplicação, estão estruturados de acordo com uma arquitetura bipartida: clientes podem se conectar a diferentes servidores e vice-versa, como mostra a figura 3.

Apesar, ou talvez por causa, da simplicidade deste modelo ele é utilizado por várias aplicações na Internet. Serviços como FTP, WWW, correio eletrônico e telnet o adotam. Em geral, nestes casos é fácil fragmentar os domínios das aplicações em sub-domínios independentes.

Note que nem sempre a pureza do modelo pode ser encontrada. Algumas vezes o modelo é levemente alterado para atender a certos requisitos não-funcionais, como, por exemplo, para melhorar a confiabilidade da aplicação.

3.1.2 Sistemas Distribuídos

Quando, para obter uma maior eficiência, confiabilidade e escalabilidade da aplicação é necessária uma maior integração entre os servidores, estes formarão um **sistema distribuído.** A cada ser-

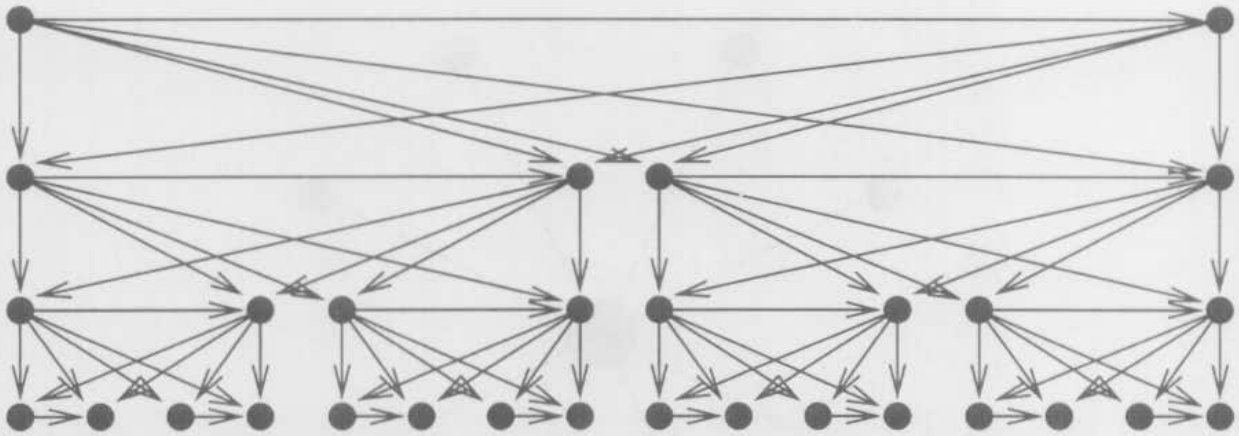

Figura 4: Esquema do grafo de delegação de zonas no DNS

vidor será delegado a responsabilidade por um sub-domínio particular da aplicação; cada s continua sendo, entretanto, administrado de forma autônoma. Na Internet o exemplo clás de um sistema distribuído seria o DNS *(Domain Name System)* [Sta87, Lot87, Moc87a, Moc87b, AL97, VDK97].

A comunicação entre os servidores, em um sistema distribuído, dá-se através de uma rede virtual de comunicação. Pode-se imaginar várias arquiteturas para esta rede. No caso do DNS, que é um sistema hierárquico, foi adotado uma espécie de árvore engordada, mostrada na figura 4. Esta arquitetura atende aos requisitos de eficiência, escalabilidade, confiabilidade e facilidade de administração necessários em um serviço essencial para o funcionamento da Internet. Uma característica importante do uso do DNS, que é a tendência das consultas ao serviço serem majoritariamente locais, favorece a estruturação quase-hierárquica da árvore engordada.

3.2 Sistemas Replicados

Quando a divisão dos domínios da aplicação não é possível, ou desejável, a alternativa é a replicação por completo da aplicação obtendo-se um **sistema replicado.** Neste caso também é necessário que o conjunto de servidores da aplicação estejam integrados para trabalharem de forma coordenada. É possível, ainda assim, que eles sejam administrados de forma autônoma, mas sujeitos a certas normas básicas de controle dependentes da aplicação.

Várias aplicações da Internet são formadas por sistemas replicados: Usenet *(net news)* [KL86, Hau92], IRC *Internet Relay Chat* [OR93], listas de discussão, redes de *mirrors* de FTP (como o CPAN [CPA], Tucows [Tuc] e o Sun SITE [Sun]) e o sistemas de caches do WWW [WC96, GS97, Obr94].

A árvore, que é a estrutura mínima em que se tem a conexidade dos servidores, é o tipo de arquitetura que aparece com mais freqüência nestes sistemas replicados. Apesar da sua vulnerabilidade à falhas dos servidores e eventuais particionamentos do sistema ela é de administração e configuração simples. Além disto a ausência de ciclos diminui a complexidade da implementação do processo de replicação. No caso da Internet, cuja arquitetura física é uma quase-hierarquia, a geração da árvore mais apropriada para a aplicação é quase direta.

A arquitetura irregular também é utilizada em sistemas replicados, como é o caso da Usenet. A presença de ciclos complica a implementação do processo de replicação mas reduz a possibilidade de degradação significativa do sistema em caso da falha de algum servidor. A Usenet é altamente dinâmica e um eventual particionamento da mesma seria muito prejudicial aos seus usuários. Como

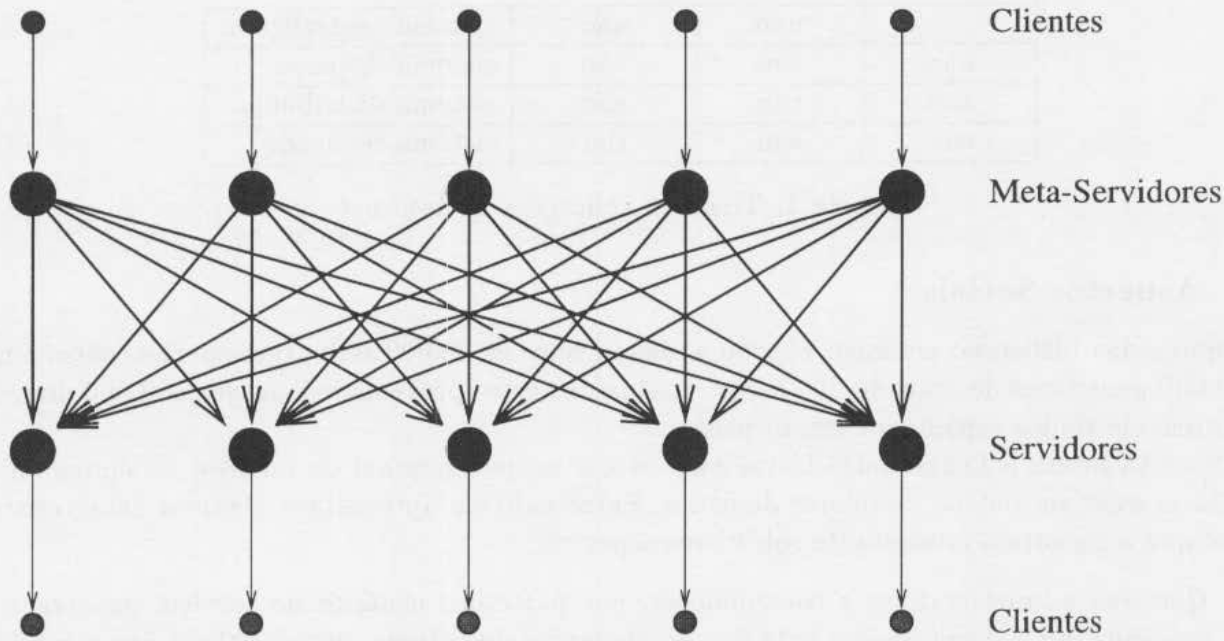

Figura 5: Esquema da arquitetura dos servidores da bras-net

a Usenet não pressupõe uma arquitetura particular a sua realização sobre a Internet pode ser efetuada de modo a otimizar o uso dos recursos da rede. Entretanto a arquitetura da Usenet reflete muito mais a arquitetura da teia social dos seus administradores *(newmasters)* do que propriamente uma busca de otimização do uso dos recursos materiais da Internet.

3.2.1 Arquitetura da Bras-net

Uma terceira arquitetura, engenhosa, é utilizada pelos servidores da bras-net[1]. À primeira vista ela parece uma estrela que teve o seu nó central replicado (uma vez para cada ponta), como mostra a figura 5. Cada ponta representa uma sub-lista abrangendo uma parcela significativa de usuários. As replicações do nó central da estrela funcionam como meta-listas. Qual a vantagem desta arquitetura em relação à estrela? Confiabilidade, carga de trabalho melhor balanceada e redução do tráfego intercontinental de mensagens.

A tabela 1 resume as definições contidas nesta seção.

4 A Esquina das Listas

A Esquina das Listas pode ser enfocada de vários ângulos: como uma comunidade brasileira emergente na Internet, como um gerenciador de listas, como a única instalação deste gerenciador de listas (no *host* dcc.unicamp.br), etc.

[1] A mais antiga lista brasileira, criada em fins de 1984, por estudantes e pesquisadores brasileiros radicados no exterior, principalmente Estados Unidos. Uma das vertentes da bras-net aflorou durante a realização, no Canadá, de um simpósio da ACM sobre sistemas distribuídos, quando Rogério Drummond, Calton Pu e outros participantes brasileiros trocaram listas pessoais de endereços conhecidos. No início a bras-net (tupinet foi outro nome sugerido) era operada "manualmente". O sucesso repentino da lista levou ao uso de um servidor de listas, abrigado na UCLA. Nos primeiros anos a lista foi administrada por Frank Schaffa[Dru98].

integração	replicação		tipo do sistema
	servidores	informação	
	não	não	sistema centralizado
não	sim	não	sistema disperso
sim	sim	não	sistema distribuído
sim	sim	sim	sistema replicado

Tabela 1: Tipos de aplicações na Internet

4.1 Aspectos Sociais

A Esquina-das-Listas, no seu auge, chegou a possuir mais de 40.000 assinaturas pertencentes a mais de 12.000 assinantes de mais de 200 listas. Assinantes estes provenientes de quase 2.000 domínios de correio eletrônico espalhados em 40 países.

De certa forma a Esquina-das-Listas não foi um projeto original ou inédito; na época da sua criação já existiam outros servidores de listas. Entretanto ela apresentava algumas características únicas que a tornaram interessante sob vários aspectos:

- Questões administrativas e contribuições, por parte dos usuários do serviço, passavam por um único *alias*, facilitando a vida do administrador de sistema, que não tinha que criar novos *aliases* para cada lista que surgisse.
- Os próprios usuários podiam criar listas nos tópicos que os interessavam.
- As instruções e os comandos para o uso da lista eram em português, o que facilitava a vida dos usuários não fluentes em inglês, língua utilizada nos outros servidores de lista.
- Havia uma interface para as listas no WWW, permitindo aos usuários eventuais participar das discussões sem se preocupar com a possibilidade de entupimento das suas caixas-postais.

Estas características explicam a popularidade atingida pela Esquina-das-Listas e a repercussão do serviço na Internet brasileira. A revista Internet World, edição brasileira, por exemplo, dedicou um largo espaço para publicar uma matéria cobrindo o projeto [Cha96].

A popularidade alcançada pela Esquina das Listas teve uma repercussão importante do ponto de vista técnico. O serviço passou a exigir uma carga de trabalho e um tráfego que superava as disponibilidades locais da sua instalação. O desafio então foi procurar soluções não convencionais que permitissem a manutenção do serviço. Desta forma a Esquina das Listas transformou-se em um laboratório real para a busca e experimentação de novas soluções técnicas relacionadas com o transporte de listas.

4.2 Arquitetura da Esquina das Listas

O serviço de lista de distribuição na Internet possui, basicamente, a arquitetura de um sistema disperso, mostrada na figura 3. Dessa forma o trabalho de transporte de listas fica repartido entre vários servidores autônomos e independentes.

Cada servidor de lista, por sua vez, que gerencia e distribui uma certa coleção de listas, é implementado segundo uma arquitetura centralizada, como na figura 2. Ou seja, pela própria natureza desta arquitetura, o trabalho de distribuição de mensagens fica bastante concentrado no servidor. Para distribuir m mensagens para n clientes o servidor teria que abrir $n \times m$ conexões com seus clientes enviando um total de $n \times m$ cópias das m mensagens.

A figura 2, entretanto, não traduz com fidelidade o que realmente acontece em um servidor de listas. A relação cliente-servidor aqui é intermediada por outro conjunto de servidores: os servidores de correio-eletrônico [Pos82]. Na verdade, vários clientes podem ter as suas caixas postais mantidas pelo mesmo servidor de correio eletrônico. Isto permite ao servidor de listas um certa economia na distribuição das mensagens: basta enviar uma cópia, destinada aos usuários de um mesmo servidor de correio eletrônico, com o cuidado de incluir, no envelope da mesma, a relação de seus recipientes.

Uma característica do serviço de listas, porém, é que seus clientes se espalham de forma bastante esparsa pela Internet. Deste modo a otimização apresentada acima tem um resultado menor do que inicialmente se espera.

É importante frisar, no entanto, que um servidor de correio eletrônico na Internet é um sistema sofisticado e que oferece várias facilidades. Uma funcionalidade que pode estar configurada em um sistema é a de permitir *relays:* a entrega de uma mensagem destinada a um domínio de correio eletrônico pode ser delegada a um servidor intermediário de correio eletrônico.

Esta funcionalidade permite a um servidor de listas a utilização dos servidores de correio eletrônico em modo solidário: os servidores de correio eletrônico podem funcionar não apenas como intermediários terminais na distribuição de mensagens mas podem colaborar repassando mensagens para outros servidores[2].

A participação solidária dos servidores permite uma redução na carga de trabalho do servidor de listas. Para distribuir n mensagens para os clientes de n servidores de correio eletrônico, por exemplo, ele poderá mandar cópias de mensagens diferentes para servidores de correio eletrônico diferentes. Cada servidor de correio eletrônico, por sua vez, terá que mandar $n - 1$ cópias de uma mesma mensagem para os outros $n - 1$ servidores participantes do sistema.

Esta solução apresenta duas vantagens: a primeira, mencionada acima, é uma melhor divisão da da carga de trabalho entre o servidor de lista e os servidores de correio eletrônico dos clientes. Além disto, esta solução permite o aproveitamento do potencial de paralelismo do sistema: enquanto o servidor de listas está enviando uma mensagem para um servidor de correio eletrônico, o servidor atendido anteriormente poderá, simultaneamente, estar repassando para seus parceiros a mensagem que ele recebeu.

Esta solução pode, e deve, ser melhorada levando-se em conta a arquitetura quase-hierárquica da Internet. Aqui os servidores de correio eletrônico são agrupados em zonas, conforme a sua proximidade relativa: servidores mais distantes do servidor de lista ficarão em zonas cujas distâncias entre si também podem ser maiores. Por exemplo, no caso da Esquina das Listas, que tinha seu servidor em Campinas, servidores da Europa foram agrupados em uma zona, servidores da Argentina em outra, servidores do Nordeste em outra, assim como os servidores do Rio de Janeiro, etc.

Esta solução leva a uma menor envergadura média nas conexões entre os servidores resultando em melhoras de latência, tráfego em linhas de longa distância e em conseqüente redução adicional da carga de trabalho dos servidores. Esta solução arquitetônica, adotada na Esquina das Listas, está esquematizada na figura 6.

Note que nesta solução o servidor de lista tem o controle total sobre o sistema. As mensagens delegadas aos servidores de correio eletrônico são enviadas com um envelope, preparado pelo servidor de lista, com a relação dos clientes destinatários daquela mensagem. É claro que nesta relação deve estar pelo menos um usuário do servidor.

[2]Esta facilidade só deve ser usada com a anuência da autoridade administrativa do servidor de correio eletrônico. Hoje ela está sendo considerada um *bug* por muitos, tendo em vista exploração abusiva que dela tem sido feito por *spamers*.

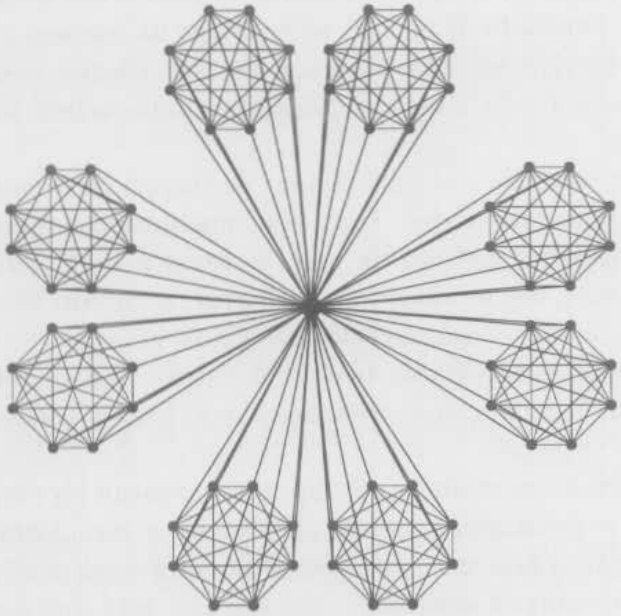

Figura 6: Arquitetura da Esquina das Listas

4.3 Escalonamento na Esquina das Listas

Ao contrário de outros servidores de listas, que delegam o trabalho de transporte de mensagens para o seu servidor de correio eletrônico local ao seu domínio, com um mínimo de preparação e organização, o servidor de listas da Esquina das Listas tem o controle completo do processo. Isto oferece a oportunidade para a introdução de melhorias no mesmo.

O escalonamento do processo de transporte de listas é organizado pelo servidor da Esquina das Listas em quatro níveis: zona, domínio de correio eletrônico, mensagens e assinantes. Um método de escalonamento diferente é utilizado em cada nível.

4.3.1 Zona

O escalonamento no nível de zona se deve a restrições de capacidade de processamento e quantidade máxima de processos em execução estabelecida pelo sistema operacional Solaris em que a esquina reside. Por isso é impossível manter conexões permanentes, simultâneas e paralelas para toda zonas.

A primeira opção é utilizar um algoritmo *round-robin:* atende-se as zonas uma após a outra em um seqüência pre-determinada. Esta é uma solução fácil de implementar; além disto ela garante que todas as zonas vão ser atendidas em algum momento, evitando a possibilidade de *starvation.* Ela apresenta desvantagens, entretanto. As zonas para as quais a transferências de listas é mais lenta podem acabar obstruindo o atendimento de outras mais rápidas.

O problema foi minimizado pela priorização do atendimento às zonas de onde provieram as contribuições mais recentes. Como existe uma certa correlação entre o número de contribuições enviadas de uma zona e sua acessibilidade a heurística tende a priorizar o atendimento às zonas mais rápidas.

Para evitar um excesso de atenção a zonas mais eficazes e reduzir as tentativas de conexões frustradas a zonas ineficazes foi implementado um sistema de temporização no atendimento às zonas. O intervalo de tempo entre atendimentos varia de um valor mínimo a um valor máximo. Toda vez que uma conexão a uma zona é frustrada o intervalo de atendimento da mesma é dobrado;

ele volta ao valor mínimo depois de atingir o valor máximo. Note, entretanto, que, independente de quando tenha sido efetuada a última remessa a uma zona, ela será atendida prontamente após o recebimento de uma nova contribuição originária da mesma (e este evento também serve para reinicializar o processo de escalonamento global relativo à fila de zonas).

4.3.2 Mensagens

No momento de atendimento de uma zona poderá haver um lote de mensagens pendentes a ser entregue. Como as mensagens são enviadas uma por vez, seqüencialmente, faz-se necessário estabelecer uma ordem de entrega.

Normalmente a ordem ideal seria a baseada em um critério temporal de antiguidade, o que satisfaria também um critério de "justiça" além de evitar *starvation* de mensagens (ou seja, toda mensagem seria entregue em algum momento).

No entanto a ordem temporal, dentro de um sistema de listas, nem sempre é a mais adequada: a atualidade pode ser mais importante do que a manutenção da ordem. Além disto, ela pode ser "injusta" sob outros ponto de vista: uma mensagem muito longa para apenas um usuário deveria realmente ser entregue antes de uma curta para vários?

A solução implementada emprega a mesma heurística utilizada no escalonamento das zonas e é contrária à ordem temporal. Mensagens mais atuais são enviadas primeiro. Ou seja, mensagens muito grandes podem acabar levando mais tempo para serem entregues (de modo a, talvez, desestimular os usuários a enviá-las); sua entrega poderá acontecer de forma parcial: só se completando para algumas zonas.

4.3.3 Domínio de Correio Eletrônico

Uma zona pode ser composta de vários domínios de correio eletrônico e conexões são feitas, efetivamente, aos servidores de correio eletrônico desses domínios. Se uma mensagem tem destinatários em vários domínios de correio eletrônico de uma mesma zona a questão é saber a qual servidor caberá a responsabilidade pela distribuição da mensagem para aquela zona.

Em primeiro lugar, se a mensagem provém da zona, o servidor escolhido será o do domínio em que se originou a mensagem. Com isto, o remetente da mesma poderá ter uma cópia da sua contribuição o mais breve possível, reduzindo uma possível ansiedade com relação à possibilidade da mensagem ter se perdido e conseqüente reenvio da mesma.

Quando a mensagem é de outra zona a eleição do servidor é definida de acordo com um sistema de contabilidade. Cada servidor tem a sua conta de deveres e haveres. O servidor cujo débito corrente é o maior será convocado para fazer a distribuição.

Débitos e créditos são computados com base em um fórmula que tenta refletir o trabalho de distribuição das mensagens. São considerados o tamanho da mensagem, o número de domínios a que ela se destina, a distância léxica entre o nome domínio do servidor convocado para a distribuição e os nomes dos outros domínios[3] e o número de usuários, naquela zona, recipientes da mensagem.

4.3.4 Assinantes

Quando a transferência de uma mensagem é delegada a um servidor, deve ser anexado ao envelope da mensagem a relação dos *endereços* dos destinatários da mesma. A Esquina das Listas ordena esta relação de modo a tentar reduzir o consumo de recursos do servidor de correio eletrônico

[3] A premissa aqui é de que há uma boa correlação entre a distância léxica e a distância efetiva entre os domínios.

invocado para distribuí-la na sua zona. É possível, entretanto, que este servidor não respeite esta ordenação para determinar a distribuição da mensagem. Para evitar que o servidor de *relay* tenha que fazer várias conexões para um mesmo domínio fim de correio eletrônico, os assinantes do domínio aparecem juntos na relação.

Muitos servidores de correio eletrônico funcionam da seguinte forma. Elas vão seguindo a ordem de recipientes no envelope e tentando conexões com os servidores de destino. Se a conexão é bem sucedida a mensagem é enviada e endereços correspondentes removidos do envelope correspondente. No caso de fracasso mensagem e envelope voltam para a fila de mensagens, onde estarão consumindo recursos de armazenamento. O servidor passa a tratar da mensagem seguinte. Ou seja, é interessante que a ordem em que os assinantes são colocados no envelope leve em consideração também a necessidade de reduzir o consumo de recursos de armazenamento no servidor de *relay*.

O esquema escolhido baseia-se também no conceito de distância léxica, mencionado acima, presumindo-se que quanto mais próximo um servidor de correio eletrônico do outro maior as chances de sucesso e mais rápidas as conexões.

5 A Proposta

Até que ponto a solução arquitetônica utilizada na Esquina das Listas poderia influenciar a construção de uma arcabouço para o transporte de listas na Internet?

Uma primeira alternativa seria a sugerir a adoção do modelo da Esquina das Listas por outros servidores de lista. Esta alternativa possui algumas desvantagens: em primeiro lugar o abuso dos servidores de correio eletrônico pela prática de *SPAM* está levando os administradores de vários domínios a não mais permitir o *relay* de mensagens pelos seus servidores. Alguns, mais radicais, consideram inclusive altamente irresponsável, da parte do administrador do sistema, a manutenção da configuração do servidor no modo promíscuo, permitindo o *relay* irrestrito.

Também pode acontecer do administrador não concordar/entender as vantagens gerais introduzidas pelo método da Esquina das Listas em relação aos métodos convencionais. Afinal este esquema aumenta a carga de trabalho dos servidores dos domínios participantes e a eventual melhoria global do sistema não traz vantagens diretas para o administrador (pode trazer até mais trabalho para ele, ao exigir sua intervenção manual para lidar com *bounces)*, apenas para os seus usuários. Isto causa um problema político cuja solução é complicada.

Um terceiro problema é que a Esquina das Listas só descentralizou a distribuição das mensagens. O gerenciamento das listas continua tendo que ser feito de uma forma centralizada. Por isto, quando delegando uma mensagem para um servidor de correio eletrônico, é necessário anexar no envelope da mesma o rol dos assinantes para os quais ela se destina. Esta informação de controle acaba representando uma parcela considerável na quantidade total de informação transferida. Quanto menor esta parcela maior seria a disponibilidade líquida de tempo para se transferir o conteúdo das mensagens, que é o que efetivamente interessa aos assinantes das listas.

Os dois primeiros problemas poderiam ser resolvidos com a modificação do protocolo de correio eletrônico de modo a permitir que servidores autentiquem seus clientes. É possível que isto venha a acontecer, principalmente como resposta ao crescente problema causado por *spams*. Um outra alternativa seria ter os servidores de listas colaborando um com o outro, seguindo um protocolo como o *Distribute* do Eric Thomas [Tho93].

Entretanto permanece o terceiro problema, que não afeta somente o desempenho e a otimização do uso da rede, mas que também causa problemas de gerenciamento e controle. Este tipo de problema não aparece na Usenet. Entretanto a Usenet só é adequada para grupos maiores e mais

densos. No caso de grupos menores e esparsos o uso da solução da Usenet implicaria em uma menor eficácia do uso da mesma e/ou em grandes problemas de gerenciamento.

A proposta de um arcabouço para o transporte de lista, a ser esboçada neste artigo pretende atender aos seguintes critérios:

- descentralização,
- eficiência,
- escalabilidade e
- flexibilidade.

Para atender a estes quatro critérios a proposta se inspira nas seguintes idéias: arquitetura do hipercubo, que permite a extensão do sistema recursivo de distribuição utilizado na Esquina das Listas e na estrutura descentralizada, distribuída e hierárquica do DNS [AL97].

5.1 O Hipercubo

O hipercubo é um tipo de arquitetura bastante utilizada em sistemas paralelos. Ele possui um série de propriedades interessantes, relacionada principalmente com sua regularidade recursiva. A rede de interconexão da *Conection Machine*, por exemplo, tem por base o hipercubo [Hil89].

Um hipercubo de ordem k e base 2 tem $n = 2^k$ vértices. Todo vértice tem grau k, que é também o diâmetro deste hipercubo.

A construção de um hipercubo de ordem $k+1$, $k \geq 0$ e base 2 é feita a partir da interligação dos vértices correspondentes de 2 hipercubos de ordem k. Um hipercubo de ordem 0, independente da base, é, por convenção, apenas um vértice isolado. A figura 7 mostra um hipercubo de ordem 4 e base 2, com dezesseis vértices, portanto. A construção de hipercubos com base maiores que 2 segue raciocínio similar. A diferença agora é que um hipercubo de ordem $k+1$ e base b deve ser construído a partir de b hipercubos de ordem k e base b.

Quando o hipercubo é usado como base para a rede de interconexão de um sistema paralelo é interessante identificar cada um dos seus vértices por um número. A numeração dos vértices do hipercubo segue um esquema parecido com a construção do mesmo. Primeiro numera-se, de forma independente, cada um dos hipercubos de ordem k que formam o hipercubo de ordem $k+1$. A seguir é só prefixar, com um dígito (entre 0 e $b-1$, onde b é a base do hipercubo) diferente para cada hipercubo de ordem k, os números de todos os vértices de cada hipercubo de ordem k.

Dada esta numeração é fácil rotear um pacote, de acordo com número/endereço de destino do mesmo. Como o endereço de dois nós vizinhos tem exatamente um dígito diferente, a distância entre dois nós pode ser determinada contando quantos dígitos diferentes tem seus endereços. O próximo nó da rota deve ser escolhido entre os vizinhos do nó corrente que estão mais próximos do nó de destino que o nó corrente. Existe pelo menos um, a não ser que o nó corrente seja ele próprio o nó de destino. No caso de um hipercubo binário de ordem k um pacote deve percorrer no máximo $k = \log n$ arestas até chegar ao nó destino.

Um outro problema que o hipercubo permite uma solução eficiente e elegante tem a ver com a difusão de um pacote por todos os nós do sistema, ou **broadcasting**. Mesmo considerando um modelo computacional bem simples, que não permite a comunicação simultânea de um nó com todos os seus vizinhos, o hipercubo permite uma solução satisfatória. O broadcasting pode ser efetuado em $\log n$ iterações; o mínimo necessário de $n-1$ mensagens circulam pelo hipercubo, e nenhum nó é sobrecarregado.

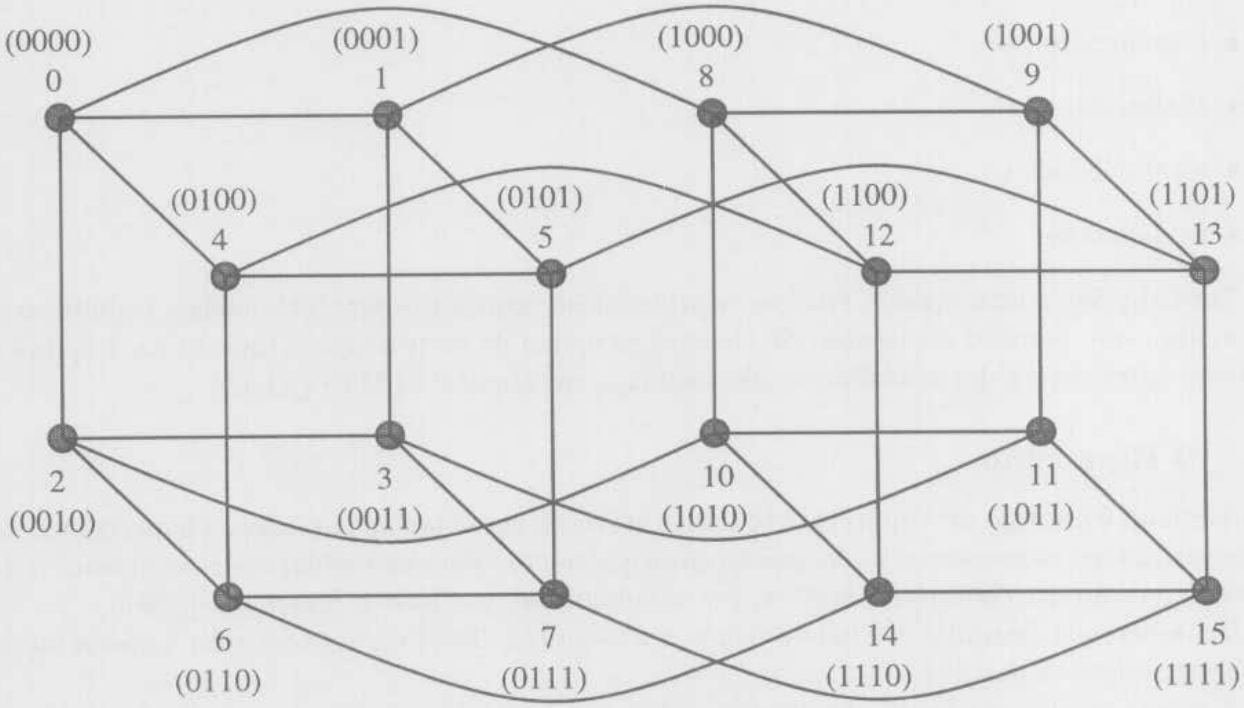

Figura 7: Um hipercubo binário de ordem 4

O algoritmo para implementação do broadcasting não é muito difícil e será omitido. A figura 8 mostra a árvore de difusão de um pacote a partir do nó 5.

5.2 O Arcabouço

O arcabouço proposto para o transporte de listas adota um sistema hierárquico. A Internet, para efeito de distribuição de listas, seria dividida em vários domínios de distribuição. No nível mais alto, cada domínio poderia corresponder a uma AS (região autônoma) ou a um continente, por exemplo.

Estes domínios de distribuição, por sua vez, poderiam ser divididos em sub-domínios e assim sucessivamente. O esquema seria semelhante ao adotado pelo DNS mas poderia ser feito de modo a melhor refletir a arquitetura quase-hierárquica da Internet.

O funcionamento deste arcabouço exigiria a presença de uma sistema de nomes como o DNS. Na verdade é possível até que, com algumas poucas extensões, o próprio DNS poderia atender as necessidades do arcabouço com relação à resolução de nomes.

Dentro desta proposta, cada domínio de distribuição possui um servidor de listas. Um servidor poderia cobrir mais do que um domínio, inclusive de níveis diferentes. O servidor ficaria responsável pela distribuição e gerenciamento das listas para os usuários do seu domínio.

As contribuições dos usuários de uma lista seriam submetidas através do próprio servidor do seu domínio. Este faria a distribuição da mensagem de forma similar à implementação de *multicasting* em um hipercubo. Ou seja, através do servidor de nomes (talvez o DNS), o servidor de listas obteria o endereço de servidores de lista (para a lista para a qual a mensagem estaria sendo submetida) dos domínios de alto nível. Estes seriam servidores quaisquer, aleatórios, que atenderiam algum sub-domínio folha dos domínios de alto nível; sem contar o próprio domínio de alto nível em que estaria o servidor de listas originário da mensagem.

O servidor então enviaria cópias da mensagens para cada um daqueles servidores, solicitando a

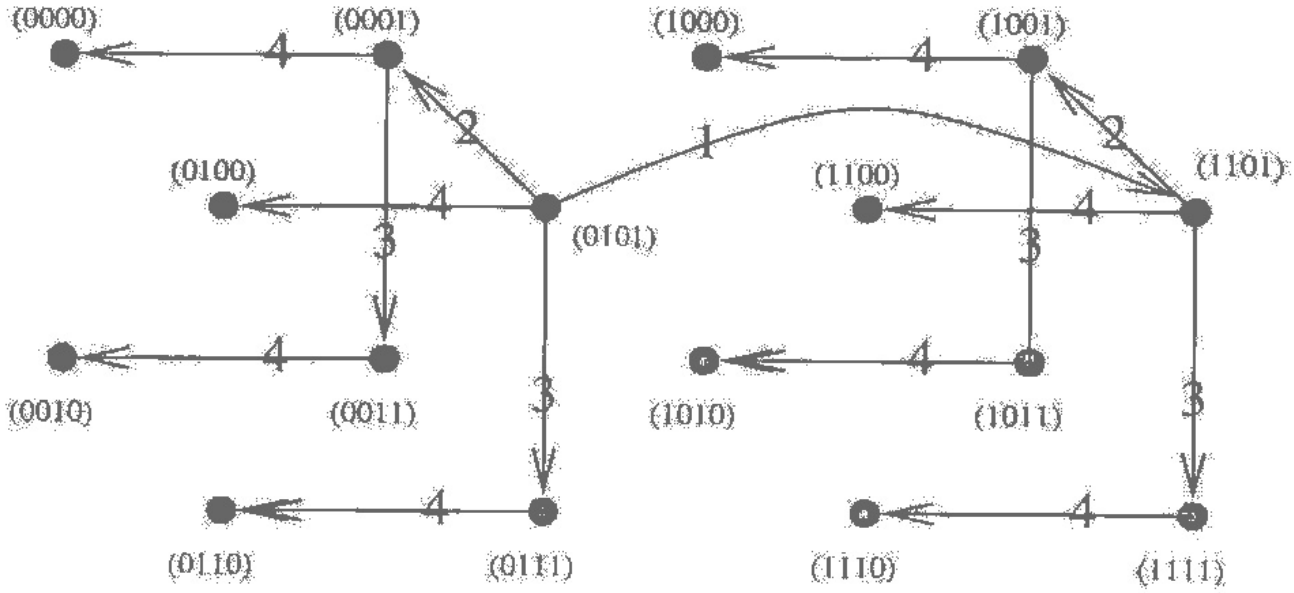

Figura 8: Árvore de difusão gerada pela disseminação de um pacote a partir do nó 5

eles a tarefa de fazer a distribuição dentro dos seus respectivos domínios de alto nível. O processo se repetiria para os sub-domínios do domínio de alto nível no qual o servidor de listas residisse e assim recursivamente. Para evitar consultas excessivas ao servidor de nomes, cada servidor de listas poderia ter o seu cache particular de endereços de servidores de listas remotos.

5.3 Discussão

O algoritmo de disseminação apresentado funciona bem em um hipercubo, mas será que este também seria o caso na Internet?

Uma análise mais realista do desempenho do algoritmo na Internet talvez devesse fazer uso de um modelo mais hierárquico tal como a árvore binária com *layout* em H da figura 9. O hipercubo poderia ser mapeado numa árvore H (vértices de hipercubo sendo mapeados nas folhas da árvore H) de mesmo tamanho e o algoritmo de difusão, acima, poderia ser analisado e implementado na mesma.

Um rastreio do algoritmo de difusão está apresentado na figura 10. Os custos da difusão, por este algoritmo, numa árvore H parecem aceitáveis e o mapeamento do hipercubo na árvore H incorre em um multiplicador igual a 2 [Sto98], que é um resultado bastante razoável e que sugere a conveniência deste algoritmo para aplicações dispersas e esparsas.

É preciso analisar outras métricas, como atraso, tráfego global a ser gerado, etc, mas a arquitetura hierárquica da árvore H, a grosso modo próxima da arquitetura da Internet, sugere que a proposta de arcabouço, apresentada neste artigo, pode ser apropriada. E não apenas para a transporte de listas, mas também, eventualmente, para outras aplicações em que se tem uma cobertura esparsa e sem muitas exigências do ponto de vista de atraso, confiabilidade e ordenação de mensagens.

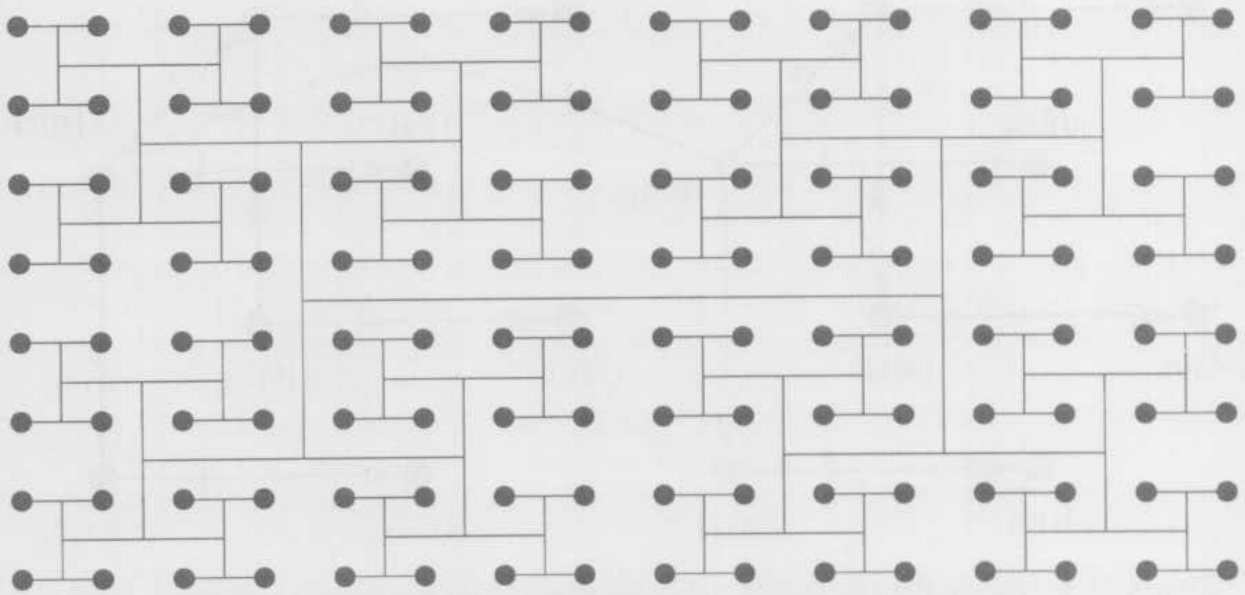

Figura 9: Uma árvore H de ordem 7

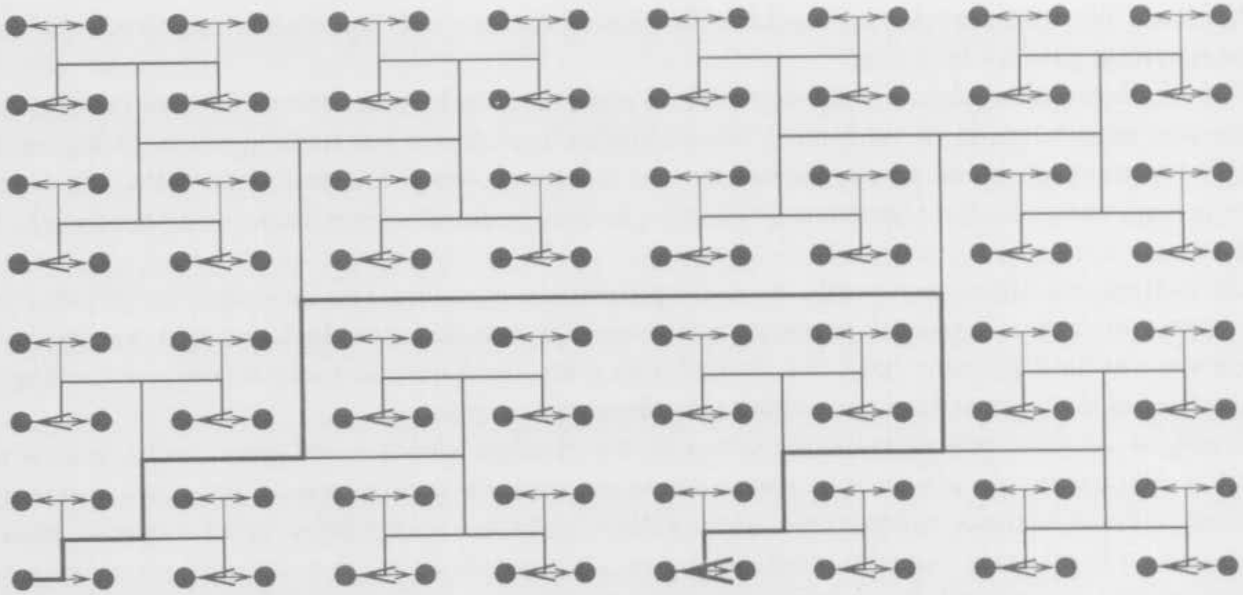

Figura 10: Árvore de difusão em um árvore H

6 Conclusão

Neste artigo foi apresentado um relato das experiências obtidas com a implantação de um serviço de distribuição de listas, a Esquina das Listas. A implantação do serviço, que parecia fácil e sem maiores conseqüências, acabou se transformando numa interessante e desafiadora tarefa. O projeto converteu-se em um laboratório bastante produtivo, permitindo a realização de experiências técnicas, sociais e políticas que só um sistema real é capaz de proporcionar. Ao longo do desenvolvimento da Esquina das Listas, surgiu a idéia de generalizar as soluções adotadas no projeto no sentido de possibilitar a construção de um arcabouço para o suporte a *multicasting* para certas aplicações distribuídas esparsas e dispersas pela Internet.

Existe uma vasta gama de alternativas a serem investigadas futuramente. A principal delas parece remeter à velha discussão na área de sistemas distribuídos sobre a forma mais apropriada de estruturar um sistema em camadas. O suporte à *multicasting* deveria ser de responsabilidade da camada de redes ou seria melhor se colocado em um nível mais alto, talvez na própria camada da aplicação? Como existe uma multitude de aplicações distribuídas é possível que colocação de suporte à *multicasting* na camada de rede, exigindo o atendimento a um largo leque de requisitos muitas vezes conflitantes, não seja prático ou viável. Uma possibilidade que precisa ser melhor analisada seria a viabilidade da utilização do MBone[Kum95] no transporte de listas.

Referências

[AL97] Paul Albitz e Cricket Liu. *DNS and BIND.* O'Reilly and Associates, segunda edição, 1997.

[CDZ97] Ken Calver, Matt Doar, e Ellen W. Zegura. Modeling Internet Topology. *IEEE Communications Magazine*, junho de 1997.

[Cha96] Sérgio Charlab. Papo na esquina é com ele mesmo. *Internet World*, pp. 56–61, Agosto de 1996.

[CPA] CPAN. Comprehensive Perl Archive Network. http://www.perl.com/CPAN/README.html.

[Dru98] Rogério Drummond. Comunicação pessoal, 1998.

[GS97] James Gwertzman e Margo Seltzer. The Case for Geographical Pushcaching. In *Proceedings of the 1995 Workshop on Hot Operating Systems*, Monterey, CA, dezembro de 1997.

[Hau92] Michael Hauben. The Social Forces Behind the Development of Usenet News, dezembro de 1992. Message-ID: ¡1992Dec9.055102.2705@news.columbia.edu¿.

[Hil89] Daniel W. Hillis. *The Connection Machine.* Artificial Intelligence. The MIT Press, 1989. paperback.

[KL86] B. Kantor e P. Lapsley. Network News Transfer Protocol. Technical report, Internet Society, fevereiro de 1986. RFC-977.

[Kum95] Vinay Kumar. *MBone: Interactive Multimedia On the Internet.* Macmillan Publishing, novembro de 1995.

[Lev87] Steven P. Leviatan. Measuring Communication Structures in Parallel Architectures and Algorithms. In Dennis B. Gannon Leah H. Jamieson e Robert J. Douglass, editores, *The Characteristics of Parallel Algorithms*, pp. 101–137. The MIT Press, 1987.

[Lot87] M. Lottor. Domain Administrators Guide. Technical report, Internet Society, novembro de 1987. RFC-1033.

[Moc87a] Paul Mockapetris. Domain Names - Concepts and Facilities. Technical report, Internet Society, novembro de 1987. RFC-1034.

[Moc87b] Paul Mockapetris. Domain Names - Implementation and Specification. Technical report, Internet Society, novembro de 1987. RFC-1035.

[Obr94] Katia Obraczka. *Massively Replicating Services in Wide-Area Internetworks*. PhD thesis, University of Southern California, Los Angeles, CA, dezembro de 1994.

[OR93] J. Oikarinen e D. Reed. Internet Relay Chat Protocol. Technical report, Internet Society, maio de 1993. RFC-1459.

[Pos82] J. Postel. Simple Mail Transfer Protocol. Technical report, Internet Society, agosto de 1982. RFC-821.

[Sta87] M. Stahl. Domain Administrators Guide. Technical report, Internet Society, novembro de 1987. RFC-1032.

[Sto98] Jorge Stolfi. Comunicação pessoal, 1998.

[Sun] Sun Microsystems, Inc. Sun SITE: Sun Software, Information and Technology Exchange. http://www.sun.com/sunsite.

[Tho93] Eric Thomas. Listserv Distribute Protocol. Technical report, Internet Society, fevereiro de 1993. RFC-1429.

[Tuc] Tucows. TUCOWS World Wide Affiliate Site Locations. http://www.tucows.com.

[VDK97] Paul Vixie, Kevin J. Dunlap, e Michael J. Karels. Name Server Operations Guide for BIND. Technical report, Internet Software Consortium, 1997. bog.

[VRST88] R. Van Renesse, H. Van Staveren, e Andrew S. Tanenbaum. Performance of the World's Fastest Distributed Operating System. *Operating Systems Review*, 22:25–34, outubro de 1988.

[WC96] Duane Wessels e K. C. Claffy. A Distributed Architecture for Global WWW Cache Integration, maio de 1996. http://ircache.nlanr.net/Cache/.

[Wiz97] Networks Wizards. Internet Domain Survey, 1997. http://www.nw.com/zone/WWW/top.html.