REALIZAÇÃO DE UM SERVIÇO DE GRUPOS LIGEIROS PARA A PLATAFORMA DE COMUNICAÇÃO EM GRUPO ENSEMBLE

Hugo Miranda, Francisco Costa e Luís Rodrigues
Departamento de Informática - Faculdade de Ciências - Universidade de Lisboa
Bloco C5 – Piso 1, Campo Grande – 1700 Lisboa
Telefone e Fax: +351-1-7500084 E-mail: hmiranda@di.fc.ul.pt

Sumário
Um serviço de comunicação e filiação em grupo permite a troca de informação de modo fiável entre diversos elementos de uma aplicação distribuída. Apresenta-se a concepção e a concretização de uma versão eficiente deste serviço, designada por Grupos Ligeiros, na plataforma Ensemble. Apresenta-se também uma aplicação distribuída que ilustra o potencial do serviço.

1. O ENSEMBLE

O Ensemble[1] é uma plataforma de comunicação fiável em grupo, desenvolvida na Universidade de Cornell, usando a linguagem de programação Objective Caml (OCaml). Os seus objectivos são: 1) suportar eficazmente a execução de pilhas de protocolos avançados; 2) suportar a aplicação de métodos formais a concretizações funcionais de protocolos de comunicação distribuída; 3) permitir a especificação de cada camada com elevada independência do sistema de execução.
O Ensemble proporciona um ambiente para composição de camadas de micro-protocolos.
Camadas adjacentes comunicam através da troca de eventos (certos tipos de eventos são descendentes enquanto outros são ascendentes). Estão previstos cerca de 40 eventos com vista a suportar uma vasta gama de micro-protocolos. Cada camada processa apenas os eventos que lhe são relevantes, entregando-os depois à camada seguinte.
Ao contrário de outros modelos de pilhas de protocolos como, por exemplo, o OSI, a aplicação não reside no topo da pilha. Os micro-protocolos podem interagir directamente com a aplicação disponibilizando interfaces laterais (fig. 1). A camada genérica para envio e recepção de mensagens (Appl) é colocada sobre camadas intermédias para optimização do desempenho[2].

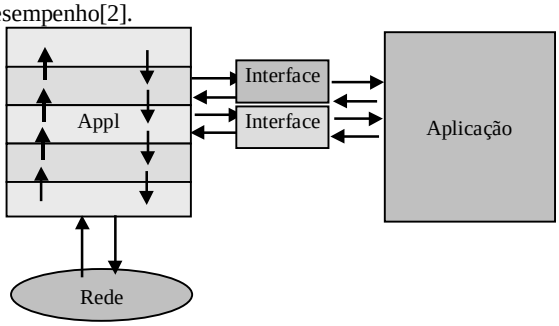

Figura 1 – Esquema de camadas do Ensemble e relação com a aplicação, onde se destaca a camada Appl para envio e recepção de mensagens.

Os interfaces são concretizados através da definição de funções (*callbacks*) que permitem o estabelecimento de comunicação bidireccional entre o Ensemble e a aplicação: o Ensemble invoca a função para indicar novos eventos, a aplicação retorna as acções que pretende executar. Um exemplo deste modelo é a funcionalidade de invocação periódica (*heartbeat*): com um período preestabelecido é invocada uma *callback*, disponibilizada pela aplicação para o efeito onde é retornado um vector com as mensagens a enviar.
A composição de interfaces permite reunir diversas funções que manipularão sequencialmente as mensagens que circulam entre o Ensemble e a aplicação, de um modo semelhante ao modelo OSI. Esta facilidade é transparente para cada um dos interfaces, sendo da responsabilidade da aplicação a sua especificação. Um exemplo desta facilidade é apresentado na figura 2.

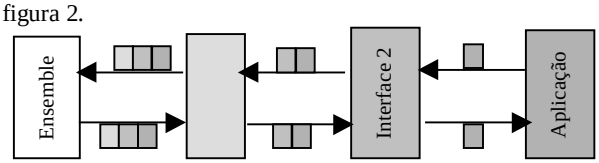

Figura 2 –Manipulação de mensagens executada pelos interfaces

A mudança de configuração da pilha em tempo de execução integra o conjunto de protocolos da plataforma. A nova configuração é proposta pela aplicação, assegurando o Ensemble a simultaneidade da sua execução por todos os membros do grupo.[2]
O Ensemble é complementado com um conjunto de serviços auxiliares, com realce para os interfaces com as linguagens C++ e Java e o serviço de CORBA.
O processo de filiação de um ponto de acesso num grupo é executado pela invocação da função de criação da pilha com os seguintes parâmetros: configuração da pilha em função das propriedades pretendidas para o serviço; nome único do ponto de acesso, opcionalmente gerado pelo Ensemble; interface(s) a utilizar e nome do grupo.
O abandono controlado de um ponto de acesso é executado através do envio de uma mensagem à pilha.

2. O SERVIÇO DE GRUPOS LIGEIROS

Um serviço de filiação e comunicação em grupo permite organizar processos em grupos, dentro dos quais são trocadas mensagens com vista a um objectivo comum. Este paradigma assegura que todos os processos de um grupo recebem informação coerente sobre os membros activos, na forma de *vistas de grupo* (*vistas*). A filiação de um grupo varia com o tempo: é possível a junção de novos processos e o abandono (voluntário ou por falhas) de outros. O serviço garante que todos os processos activos recebem a mesma sequência de vistas de grupo e, entre cada par de vistas sucessivas, exactamente as mesmas mensagens. Esta funcionalidade, designada por Sincronia Virtual[5] facilita a construção de aplicações distribuídas.
Para proporcionar este serviço é necessária a execução de detectores de falhas e de protocolos que garantam acordo e ordenação.
Naturalmente, a execução destes componentes consome recursos, como largura de banda e tempo de processamento, embora na maioria das situações o impacto no desempenho seja pequeno. No entanto, se considerarmos situações em que uma larga percentagem de processos partilha um conjunto de grupos, constata-se que existe uma redundância desnecessária das operações realizadas por esses componentes.
A partilha destes recursos pode ser alcançada projectando vários grupos de nível de utilizador num único grupo de sincronia virtual.
Os primeiros são chamados grupos ligeiros (*light-weight groups*) em contraste com os grupos de mais baixo nível a que se dá o nome de grupos “pesados” (*heavy-weight groups*). Ao serviço que faz a projecção dos grupos ligeiros em grupos pesados dá-se o nome de Serviço de Grupos Ligeiros (*light-weight group service*).
Uma das características desejáveis neste serviço é a sua transparência para a aplicação e para a plataforma de suporte.

Torna-se assim possível acrescentar uma mais-valia às aplicações já existentes e eliminar um factor adicional de complexidade na programação destes sistemas.
Este serviço foi já realizado em vários sistemas de comunicação em grupo, de que são exemplo: Delta-4[4], Isis[5] e Horus[6].
Um dos problemas inerentes ao modelo dos grupos ligeiros é o da interferência gerada pela partilha do grupo "pesado": os seus bloqueios (que podem surgir por alterações na filiação ou por solicitação de um dos membros) são repercutidos em todos os grupos ligeiros; se for usado o protocolo FIFO a perda de mensagens de um grupo atrasa o envio das restantes. Desta enumeração conclui-se que existe um equilíbrio entre as vantagens proporcionadas pela partilha de recursos e as desvantagens induzidas pela interferência. Nalguns casos, a existência de um único grupo pesado no serviço pode revelar-se prejudicial. A solução passa por estabelecer heurísticas que, baseadas na composição dos grupos ligeiros, determinem, preferencialmente em tempo de execução, projecções adequadas à topologia, ou seja, reunindo no mesmo grupo "pesado" os grupos ligeiros que contêm elementos com distribuição semelhante pelos processos[3].

3. OS GRUPOS LIGEIROS E O ENSEMBLE

O Ensemble é um sistema no qual é vantajosa a aplicação de um serviço de grupos ligeiros. Trata-se de uma plataforma aberta, concebida com o intuito de facilitar a investigação nesta área. Inclui já camadas que satisfazem a sincronia virtual, apresentando latências muito baixas, o que levou à sua utilização em diversos outros projectos[1]. Tratando-se também de uma plataforma muito recente, esta realização permite ainda avaliar as suas capacidades de expansão, nomeadamente, confrontando a modularização que apresenta com os requisitos exigidos pelo serviço.

4. CONCRETIZAÇÃO

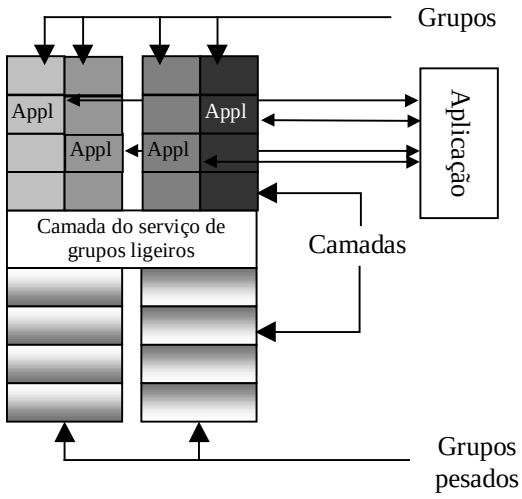

Figura 3 – Modelo previsto de realização do serviço de grupos ligeiros no Ensemble

Capacidade de configuração do Ensemble
O modelo inicialmente previsto consistia no desenvolvimento de uma camada que seria integrada na pilha de micro-protocolos (fig. 3). A camada cruzaria horizontalmente as pilhas instaladas para os diferentes grupos pesados e possibilitaria aos grupos ligeiros diferentes configurações de pilhas.
Esta concretização não foi possível devido à incapacidade do Ensemble em suportar a composição em forquilha de pilhas de micro-protocolos.
Como alternativa, a solução encontrada passou por colocar o serviço no interface entre a camada genérica Appl e a aplicação. Esta solução é esquematicamente apresentada na figura 4.

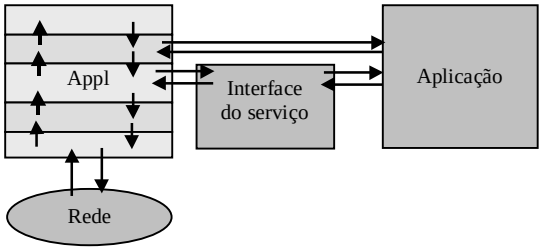

Figura 4 – Modelo utilizado na realização do serviço de grupos ligeiros no Ensemble

A colocação do serviço exteriormente à pilha de micro-protocolos levanta algumas restrições ao nível da concretização: 1) o posicionamento rígido do serviço limita as possibilidades de configuração do sistema; 2) todos os grupos ligeiros associados a um grupo pesado partilham exactamente a mesma configuração da pilha; 3) o posicionamento do serviço na arquitectura retira o acesso aos eventos que circulam na pilha e que não são passados à aplicação; 4) o suporte a múltiplos grupos "pesados" torna-se mais complexo; 5) a realização assume um carácter pouco intuitivo.
Do ponto de vista da aplicação não são, no entanto, sentidas grandes alterações. O interface do serviço é composto com os restantes, assegurando à aplicação a utilização de todas as facilidades de que anteriormente dispunha.
Funcionalidades actuais e futuras
As principais funcionalidades do serviço foram já concretizadas. É já possível a sua utilização de forma transparente para a aplicação quer no que respeita à recepção e envio de mensagens quer no processamento de eventos (por exemplo, alterações na vista do grupo). Para utilizar este serviço numa aplicação já existente será apenas necessário proceder à alteração do nome da função de criação da pilha (que mantém o conjunto de parâmetros da original) e adicionar o interface do serviço à composição de interfaces pretendida.
No futuro, o serviço será melhorado adicionando o suporte a múltiplos grupos "pesados" e possibilitando a mudança dinâmica de grupos ligeiros entre eles. A pilha relativa ao grupo pesado passará a ser especificada no código da aplicação e não no do serviço. Finalmente, o período de *heartbeat* para todos os grupos ligeiros deixará de estar condicionado ao solicitado pelo primeiro grupo ligeiro instalado.
Estruturas de dados
Cada instância do serviço, pela natureza das suas funções, lida com dois tipos de pontos de acesso: os pontos de acesso reconhecidos pelo Ensemble (pontos de acesso de grupos "pesados" ou PAP's) e os pontos de acesso da aplicação (pontos de acesso de grupos ligeiros ou PAL's). Ao serviço de grupos ligeiros cabe a responsabilidade de garantir o mapeamento correcto entre ambos, ocultando os PAP's à aplicação e os PAL's à plataforma, como apresentado na fig. 5.

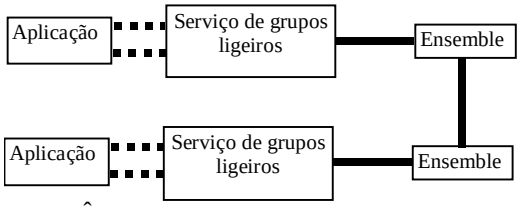

Figura 5- Âmbito dos pontos de acesso de grupos ligeiros (PAL's), a tracejado e dos pontos de acesso de grupos pesados (PAP's) a cheio

A informação contida num trio ordenado (Id do grupo ligeiro,PAL,PAP) em que o PAP suporta o PAL e este pertence ao grupo ligeiro é a estrutura base de toda a

informação necessária à execução das projecções entre o Ensemble e a aplicação.
O serviço é totalmente descentralizado, não se baseando na existência de um coordenador. Para cada grupo ligeiro a lista *vista* contém o conjunto ordenado de membros do grupo, identificados individualmente por pares ordenados da forma (PAP,PAL). Esta lista é replicada por todas as instâncias do serviço.
Cada instância mantém ainda o seu estado relativamente a todos os grupos ligeiros (fig. 6).
Como estruturas de apoio, cada grupo ligeiro suporta ainda as listas *junção* e *abandono* com o mesmo formato da lista *vista*.

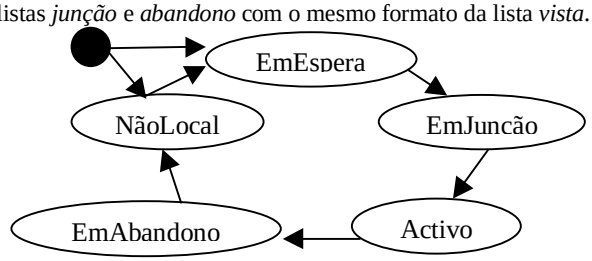

Figura 6 – Diagrama de Estados de um Grupo Ligeiro numa instância do serviço

Sincronização virtual nos grupos ligeiros
A manutenção da propriedade de sincronia virtual para os grupos ligeiros é conseguida através do grupo "pesado" de suporte. O serviço provoca mudanças de vista do grupo "pesado" sempre que tal se justifique para um dos grupos ligeiros associados. Presentemente, estas mudanças de vista são provocadas durante o processo de filiação e de abandono de pontos de acesso a um grupo ligeiro.
Protocolos
a) Encapsulamento de mensagens
Uma vez que a utilização do serviço se pretende transparente para a aplicação e para o Ensemble, é responsabilidade do serviço encapsular/desencapsular as mensagens da aplicação com vista a assegurar: 1) a circulação de mensagens coerentes na perspectiva do Ensemble, uma vez que os grupos ligeiros, conhecidos pela aplicação, não estão registados na plataforma; 2) a sua entrega aos destinatários correctos, o que implica a identificação unívoca de todas as relações PAP/PAL dos grupos ligeiros. Foi por isso criado um conjunto de estruturas de dados para troca de mensagens entre instâncias do serviço. A sua projecção nas acções executadas pela aplicação está representada na figura 7.

Estrutura de encapsulamento	**Sentido**	**Acção da Aplicação**
LSend(grupo, origem, destino[], msg)	←→	**Send**(destino[], msg)
LCast(grupo, origem, msg)	←→	**Cast**(msg)
LLeave(grupo, origem)	←	**Leave**()
LJoin(grupo, origem)		**Pedido de criação de pilha**
LState(lista(grupo, ponto))		

Figura 7 – Relação entre as estruturas de encapsulamento do serviço e as mensagens da aplicação

As primitivas LSend, LCast e LLeave contêm redundância: conhecido o PAP que enviou a mensagem (informação disponibilizada pelo Ensemble) e o grupo ligeiro em causa (identificado na mensagem) é possível encontrar, no(s) destinatário(s) o PAL de origem pela consulta à lista *vista* respectiva. No entanto, a sobrecarga introduzida pela adição do identificador *origem* à mensagem é menor que a introduzida pela pesquisa sequencial numa lista em um ou mais destinatários.
b) Filiação
Os estados *EmEspera* e *EmJunção* estão associados ao processo de filiação de um ponto local num grupo ligeiro. O primeiro quando a aplicação realizou o pedido mas a mensagem não foi ainda enviada pelo serviço, o segundo enquanto o serviço aguarda a instalação de uma nova vista onde o novo ponto será incluído. Nesse momento, a instância do serviço actualiza o estado local deste grupo ligeiro para *Activo*. De notar que a existência dos dois estados se prende apenas com o formato de *interface* do Ensemble que condiciona o envio de mensagens à resposta a *callbacks*.
Aquando da recepção da mensagem, todas as instâncias do serviço adicionam à sua lista *junção* para o grupo ligeiro o par (PAP,PAL) onde o PAP refere o ponto de acesso detido pela instância do serviço que solicitou a filiação no grupo ligeiro e o PAL o nome do ponto de acesso do grupo ligeiro. Os membros de *junção* são adicionados à lista *vista* aquando da instalação da nova vista pelo grupo "pesado".
O processo de filiação de um participante num grupo ligeiro está representado na figura 8.

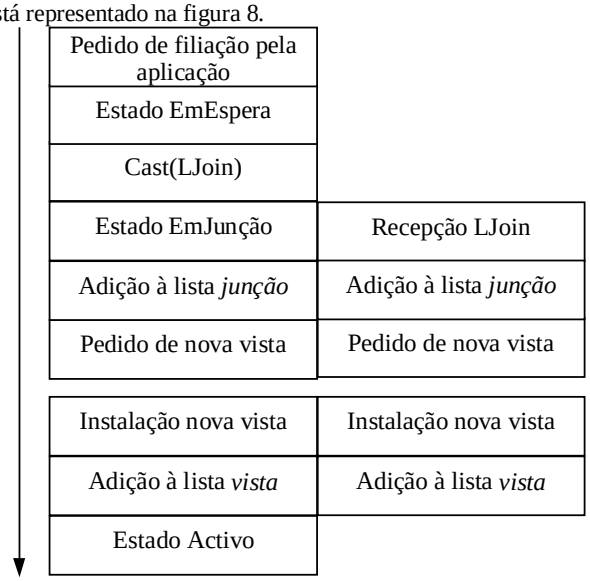

Figura 8 – Acontecimentos relevantes na filiação de um ponto a um grupo ligeiro. Pelo processo que realiza o pedido (à esquerda) e pelos restantes (à direita)

Note-se que uma eventual falha do processo em junção leva a que os restantes membros do serviço não o incluam na lista *vista*, uma vez que o ponto não consta da vista entregue pelo Ensemble.
Um caso particular do problema de junção refere-se à união de partições, que sucede aquando da filiação de um ou mais pontos de acesso a um grupo pesado. Nesse caso, o Ensemble instala inicialmente uma vista de transferência de estado que sincroniza as listas *vista* de todas as instâncias do serviço. Só após a conclusão desta vista (determinada pela plataforma) é instalada a nova vista para a aplicação.
c) Abandono
O abandono de um participante da aplicação pode dar-se de forma controlada ou por falha.
No primeiro caso a aplicação envia a mensagem de abandono disponibilizada pela plataforma (*Leave*) que é difundida pelo serviço encapsulada numa estrutura de tipo *LLeave*. Cada instância do serviço adiciona então o par (PAP,PAL) respectivo à sua lista *abandono*. Quando é instalada uma nova vista pelo Ensemble, todas as instâncias do serviço excluem o(s) ponto(s) em abandono da lista *vista* e a aplicação que desencadeou o processo recebe o evento de saída do grupo. O estado do grupo ligeiro na instância do serviço do ponto em abandono é actualizado inicialmente para *EmAbandono* e para

NãoLocal no final do processo. O ponto em abandono continua a receber mensagens até à instalação da nova vista. Assegura-se assim a propriedade da sincronia virtual.
O abandono por falha é detectado pela comparação das vistas instaladas pelo Ensemble (que contém os PAP's) com a informação contida nas listas *vista*. A existência de um PAP numa lista *vista* mas não na vista instalada pelo Ensemble representa uma situação em que o processo terminou abruptamente e portanto deve também ser excluído das vistas dos grupos ligeiros em que eventualmente estivesse filiado.
Em qualquer um dos casos mantém-se o princípio definido pela plataforma de informar a aplicação do abandono de outros membros apenas pela sua exclusão da vista instalada.

5. SERVIÇO DE LEILÕES ON-LINE

Nesta secção é apresentada uma aplicação que ilustra a utilidade do serviço de grupos ligeiros que tem vindo a ser descrito.
A aplicação proporciona um serviço de leilões, os quais são terminados após um período de tempo estabelecido antecipadamente. Este método alternativo de terminação visa: 1) atenuar as diferenças de velocidade de acesso à Internet; 2) libertar os utilizadores da atenção permanente ao serviço.
O serviço é composto por três módulos: o servidor de produtos, o leiloeiro e os licitadores. Os dois primeiros acessíveis apenas pela organização que oferece o serviço e o terceiro pelos clientes. O servidor de produtos tem como missão o armazenamento e disponibilização da informação referente aos produtos leiloáveis, bem como a coordenação do grupo de controlo, que será apresentado mais à frente. O cliente pode acompanhar o leilão de diversos produtos em simultâneo, enviando para cada um deles licitações e comentários, recebidos por todos os outros elementos e pelo leiloeiro. A janela de acompanhamento de cada produto é apresentada na figura 9.

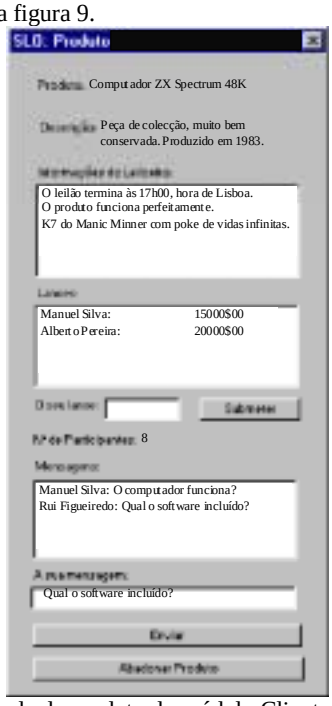

Figura 9 – Janela de produto do módulo Cliente do serviço de leilões on-line

O leiloeiro coloca os produtos em leilão. O seu interface é muito semelhante ao dos clientes, distinguindo-se o seu código por declarar o encerramento e o vencedor do leilão.
No plano da realização, a aplicação utiliza duas classes de grupos (figura 10): 1) um grupo para cada produto presentemente em leilão, onde circulam as mensagens dos clientes e do leiloeiro: licitações, mensagens do leiloeiro, mensagens dos participantes, encerramento do leilão e declaração do vencedor; 2) um grupo de controlo, coordenado pelo Servidor de Produtos que informa quais os produtos presentemente em leilão e os nomes dos grupos associados. Identificado univocamente em toda a aplicação, é o primeiro grupo em que se filiam todos os processos.

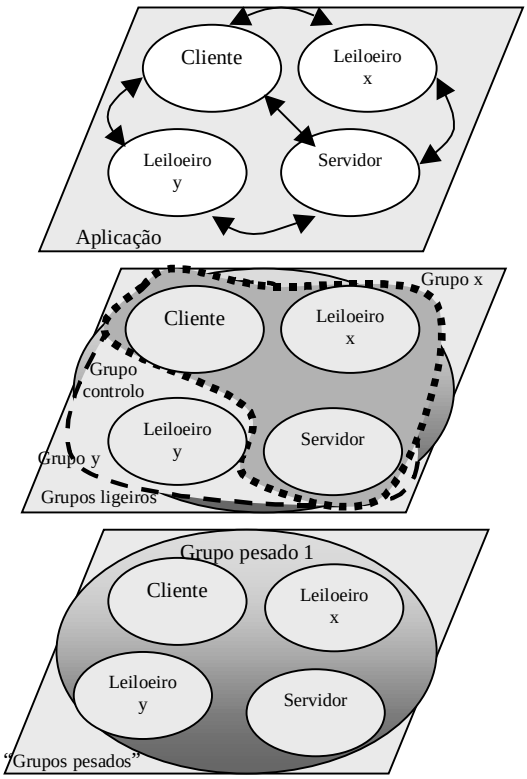

Figura 10 – Modelo da arquitectura da aplicação de leilões

A estrutura de grupos apresentada favorece a utilização do serviço de grupos ligeiros uma vez que se admite como provável que os clientes assistam a diversos leilões em simultâneo. A optimização de recursos é conseguida ao nível do servidor e dos clientes. Ambos conseguem uma redução do número de grupos pesados que subscrevem; o que irá diminuir o tempo dedicado à gestão dos próprios grupos, libertar o servidor para as actividades para que foi concebido, diminuir o tráfego na rede da organização e melhorar a qualidade de acesso do cliente, aumentando assim a sua satisfação.
O trabalho foi desenvolvido na linguagem Java utilizando o interface Ejava do Ensemble.

6. CONCLUSÕES E TRABALHO FUTURO

Tendo já sido provada a eficiência dos grupos ligeiros em ambientes onde uma grande percentagem de membros comuns subscreve vários grupos[3] realizou-se este serviço sobre uma plataforma recente que apresenta algumas características inovadoras: o Ensemble. Apresentámos também um exemplo de uma aplicação onde é vantajosa a utilização de grupos ligeiros.
A realização permitiu-nos concluir que: 1) o modelo de decomposição da pilha do Ensemble possui algumas limitações; 2) o Ensemble apresenta características inovadoras que permitiram contornar o problema realizando o serviço na zona de *interface*, externa à pilha embora com resultados não tão satisfatórios e intuitivos; 3) o paradigma funcional desempenhou um papel importante na realização, nomeadamente através da composição de funções, possibilitando a criação de um serviço transparente para a plataforma e para a aplicação.
Num futuro próximo será concluída a realização do serviço, alargando o suporte a múltiplos grupos pesados e estabelecendo critérios de mapeamento dinâmico que

optimizem o serviço. Para além do ponto anterior, a diminuição da interferência passará também pela realização de um algoritmo de bloqueio ao nível dos grupos ligeiros.

REFERÊNCIAS

[1] M. Hayden, "The Ensemble System", PhD. Thesis, Cornell University, 1998.

[2] R.van Renesse, K. Birman, M. Hayden, A. Vaysburd, D. Karr, "Building Adaptive Systems Using Ensemble", Cornell University Technical Report, TR97-1638, July 1997.

[3] L. Rodrigues, K. Guo, P. Veríssimo, K. Birman, "A Dynamic Light Weight Group Service", In *Proceedings of the 15th IEEE Symposium on Reliable Distributed Systems*, pages 130-139, Niagara-on-the-Lake, Canada, October 1996

[4] D. Powell, editor. *Delta-4 – A Generic Architecture for Dependable Distributed Computing*, Springer Verlag, November 1991.

[5] K. Birman and R. van Renesse, editors. *Reliable Distributed Computing With the ISIS Toolkit*, IEEE CS Press, March 1994.

[6] R. van Renesse, K. Birman and S. Maffeis, Horus, a flexible group communication system. *Communications of the ACM*, 39(4); 76-83, April 1996