Modelos de Distribuição para Recolha de Informação de Gestão

Paulo Simões, João Rodrigues, Luís Silva, Fernando Boavida
CISUC – Universidade de Coimbra
Pólo II, Dep. Eng. Informática, P-3030 Coimbra
{psimoes, jpr, luis, boavida}@dei.uc.pt

Resumo[*]

Neste artigo é apresentado um estudo experimental onde se compara o comportamento de diversos modelos de distribuição, baseados em agentes móveis, em operações de recolha e processamento de informação de gestão dispersa para rede. O desempenho constitui a principal métrica deste estudo, mas são também considerados outros factores, tais como o tráfego de rede e os custos de arranque associados a cada modelo. Este estudo mostra que em muitas situações o desempenho dos sistemas de agentes móveis é determinado essencialmente pela distribuição do processamento, e não pela localização dos agentes. O estudo mostra também que em geral os modelos migratórios penalizam significativamente o desempenho do sistema, ainda que possam ser relevantes noutras vertentes, tais como a flexibilidade e adaptabilidade das aplicação de gestão.

Palavras-Chave: *Gestão de Redes, Agentes Móveis*

1. Introdução

O termo "Gestão Distribuída" engloba actualmente um número bastante alargado de abordagens distintas à gestão de redes descentralizada [2], que partilham entre si a noção de que a distribuição do processo de gestão por diversos nós da rede resultará em melhores soluções de gestão, caracterizadas por níveis acrescidos de flexibilidade, eficiência e robustez. A propostas pioneiras, tais como a Gestão por Delegação [3], sucederam mais tarde diversas abordagens diferenciadas, quer no contexto da evolução das arquitecturas clássicas de gestão [4-5], quer seguindo mais de perto os avanços entretanto registados no campo da computação distribuída, aproveitando por exemplo o CORBA [6], os agentes inteligentes [7], o Java [8] e o Jini [9].

Os agentes móveis (AM) representam um dos últimos paradigmas a seguir este percurso de conversão à gestão distribuída. A principal diferença entre os AM e outras tecnologias de computação distribuída reside na capacidade de alterar dinamicamente a localização do agente durante a sua execução, sem com isso perder os seus dados ou mesmo o seu estado de execução. Daqui resultam, potencialmente, metáforas de programação mais naturais e maior capacidade de adaptação dinâmica, mesmo em ambientes adversos como computação móvel e *disconnected computing* [10].

Nos últimos anos foram conduzidos diversosestudos sobre a utilização da tecnologia de agentes móveis na área de gestão de redes. Alguns estudos analíticos [11-14] centraram-se em aspectos específicos, tais como a eficiência, escalabilidade e sensibilidade às condições da rede das soluções baseadas em AM. No entanto, apesar da qualidade destes estudos, a sua correspondência com aplicações reais está ainda por determinar. Existem ainda diversas avaliações de carácter mais experimental, em geral com o objectivo de validar implementações específicas (e.g. [15-17]) mas também, se bem que com menor frequência, abordando esta temática de forma mais abrangente [18-22].

Nesta última categoria destaca-se o trabalho desenvolvido por Gavalas et al., focado na problemática da recolha de grandes volumes de informação de gestão. Por meio de técnicas de

[*] Este artigo corresponde a uma versão revista e adaptada de trabalho anteriormente publicado [1]

compressão semântica dos dados directamente na fonte [20], de filtragem de tabelas [21] e de esquemas de distribuição hierárquica [22], este trabalho demonstra que é possível, usando agentes móveis, optimizar significativamente o tráfego de rede e o desempenho das aplicações de gestão. Ainda que o refinamento de técnicas de compressão de dados constitua o principal objectivo desta linha de trabalho, as diversas técnicas de distribuição exploradas são também bastante interessantes. No entanto, o uso de ambientes de teste com dimensões reduzidas impede a extrapolação dos resultados para outros ambientes de aplicação.

O estudo por nós conduzido complementa essa linha de trabalho e é também baseado em aplicações de recolha e tratamento periódico de grandes volumes de informação de gestão dispersa pela rede. Em cada ciclo são conduzidas diversas transacções SNMP com os nós geridos, local ou remotamente. Os dados obtidos são então processados, comprimidos e enviados para a estação de gestão central. No entanto, a utilização de um ambiente de testes de maiores dimensões e a maior atenção prestada ao comportamento do sistema em estado estacionário (isolando os custos de arranque dos custos normais de funcionamento) acabou por nos conduzir a uma nova perspectiva sobre esta temática.

Na Secção 2 deste artigo é descrito o ambiente de testes por nós usado, enquanto nas Secções 3 e 4 se discute o desempenho de modelos de distribuição baseados, respectivamente, em mobilidade restrita e em agentes migratórios. Na Secção 5 são apresentadas as medições do tráfego de rede e na Secção 6 é analisado o comportamento dos diversos modelos considerados em redes com reduzida largura de banda. Na Secção 7 são discutidos os custos de arranque e na Secção 8 são apresentadas as principais conclusões do estudo.

2. Ambiente de Testes

Foram considerados cinco modelos distintos de distribuição (Figura 1). O **modelo estático centralizado (EC)** corresponde à aplicação clássica de gestão centralizada, na qual uma estação central consulta directamente – por meio do transacções SNMP – os diversos nós geridos. No **modelo migratório (MG)**, um único agente móvel visita sucessivamente cada um dos nós geridos de modo a obter directamente a informação de gestão, por meio de transacções SNMP locais. Quando termina o seu périplo o agente envia os dados entretanto recolhidos e comprimidos para o ponto central. No **modelo migratório delegado (MD)** este processo é repartido por diversos agentes móveis que trabalham em paralelo, cada um deles visitando um conjunto distinto de nós. No **modelo *master/worker* (MW)** é enviado um agente para cada nó gerido, onde permanece estacionário conduzindo transacções SNMP locais e comprimindo os dados obtidos antes do seu envio para o ponto central. No **modelo estático delegado (ED)** a rede é dividida em vários domínios de gestão. Em cada domínio existe um único agente móvel que permanece estacionário num dos nós, conduzindo a partir daí transacções SNMP com os restantes nós do domínio e enviando depois os resultados, agregados e comprimidos, para o ponto central. Em todos os modelos as transacções SNMP com cada nó são conduzidas de modo sequencial (i.e. uma nova transacção com o nó Y só é iniciada quando termina a transacção anterior com esse mesmo nó). No entanto, nos modelos EC e ED cada gestor conduz as transacções com os diversos nós geridos de modo independente e paralelo. Todos estes modelos de distribuição são familiares: o EC representa a maioria das aplicações centralizadas baseadas em SNMP e o ED corresponde a uma simples delegação hierárquica. Os restantes modelos foram também já abordados em estudos anteriores [20-22], ainda que com designações diferentes.

O ambiente de testes consistiu em estações Intel relativamente homogéneas (Windows NT4, Pentium II 350 MHz e 128 Mbyte de memória), interligadas por uma rede *ethernet* comutada. O suporte de agentes móveis foi fornecido pela plataforma JAMES [23-24] e cada nó gerido incluiu, para além do serviço SNMP nativo do Windows NT4, uma *agência* JAMES que lhe permite receber agentes móveis.

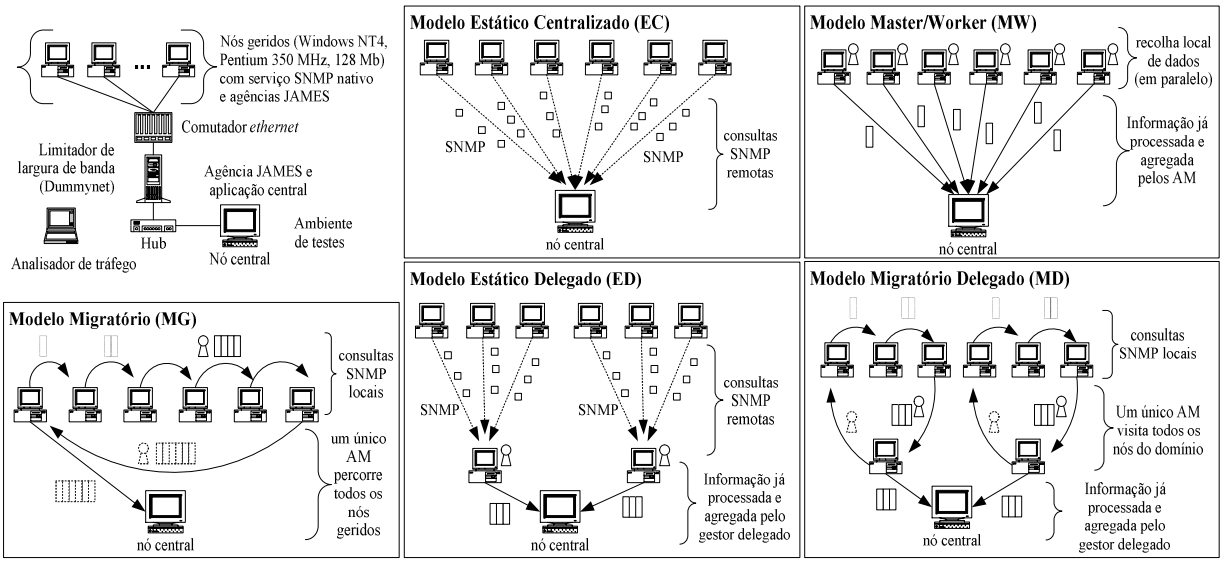


Figura 1: Ambiente de Testes e Modelos de Distribuição

Algumas medições foram efectuadas usando a rede comutada a 10 Mbps, mas condições mais restritivas foram também testadas por meio de um limitador de largura de banda [25] colocado entre o nó central e os nós geridos. Ainda que por razões práticas não tenha sido limitada a largura de banda entre os nós geridos, considera-se que o ambiente de testes não deixa por isso de ser suficientemente representativo, uma vez que nos modelos EC e MW não existe qualquer troca de dados entre nós geridos e nos modelos delegados (ED, MD) só existe comunicação entre os nós do mesmo domínio de gestão – habitualmente as fronteiras dos domínios de gestão não atravessam redes locais.

Os dados enviados para o nó central são previamente comprimidos, excepto no modelo EC. Por meio de conversão dos formatos de representação dos dados, o volume correspondente a cada "objecto" é reduzido a 20% do seu tamanho original numa mensagem SNMP. No âmbito dos testes considerou-se que este mecanismo de compressão dos dados seria representativo de técnicas de compressão mais realísticas, quer do ponto de vista do esforço computacional quer do ponto de vista das taxas de compressão obtidas.

Foram introduzidos diversos parâmetros durante os testes, tais como o número de nós geridos (entre 1 e 120), a quantidade de informação de gestão recolhida (entre 25 e 600 objectos da MIB-II e do tipo INTEGER, por nó gerido e por ciclo) e a largura de banda entre o nó central e os nós geridos (32 Kbps a 10 Mbps). Outros parâmetros também considerados nos testes, tais como técnicas adicionais de optimização do número de transacções SNMP e mecanismos de compressão mais agressivos, não serão abordados neste artigo.

As medições obtidas incluem o desempenho do sistema e o tráfego de rede medido junto ao ponto central. O desempenho do sistema é representado pelo intervalo de tempo que decorre entre o início de um novo ciclo de recolha de dados (desencadeado pelo nó central) e a recepção e tratamento, no nó central, dos dados correspondentes.

De modo a isolar os custos de arranque dos custos de funcionamento "regular", o primeiro ciclo – no qual os agentes estacionários são colocados nos nós geridos e onde os agentes migratórios não beneficiam de mecanismos de *caching* de código – e os ciclos seguintes foram medidos de modo independente. Os resultados de desempenho aqui apresentados correspondem à média de 9 ciclos "regulares", com um desvio padrão inferior a 5%. No entanto, apesar destas precauções, a natureza do ambiente de testes (dezenas de estações NT, serviços SNMP nativos e agentes móveis baseados em Java) afectou os resultados com algum ruído perceptível, em particular nos modelos mais rápidos.

3. Desempenho dos Modelos de Mobilidade Restrita

3.1 O Efeito da Localidade

De modo a avaliar o efeito da localidade no desempenho, foi efectuado um primeiro teste bastante simples, usando duas máquinas exactamente iguais (1 nó central e 1 nó gerido) e ajustando a largura de banda da ligação entre 32 Kbps e 4 Mbps. As medições experimentais mostram que o desempenho relativo do modelo MW, quando comparado directamente com o modelo EC, melhora com a progressiva restrição da largura de banda e com o aumento do volume da informação de gestão recolhida (Figura 2, gráfico à esquerda).

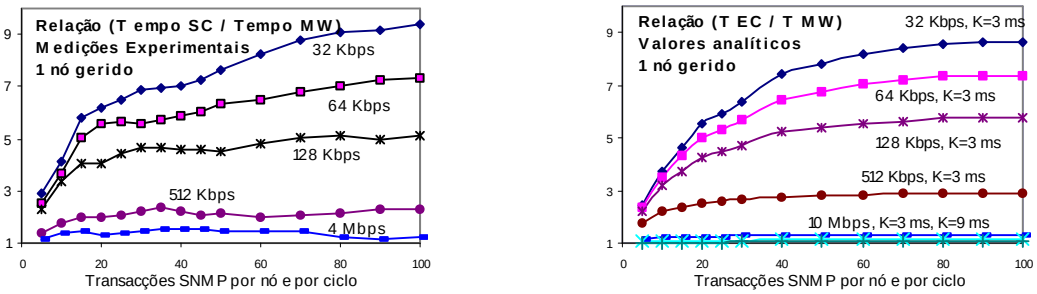


Figura 2: Relação entre o Desempenho dos Modelos EC e MW (1 nó gerido)

Contudo, este aumento de desempenho associado à localidade das transacções SNMP têm um tecto máximo, determinado pelas condições da rede, pela capacidade de processamento dos nós geridos e pela taxa de compressão da informação de gestão. Os tempos de resposta esperados para os modelos EC e MW são definidos de forma aproximada, respectivamente, pela Expressão 1 e pela Expressão 2. Nestas expressões *N* representa o número de transacções SNMP por ciclo e por nó gerido (no nosso caso cada transação envolve 5 objectos da MIB-II). $T_{getSNMP}$ e $T_{respSNMP}$ representam, respectivamente, o tempo necessário para construir um pedido SNMP e para descodificar e processar a respectiva resposta. T_{AgSNMP} representa o tempo gasto pelo serviço SNMP para receber processar e responder ao pedido. Os nossos dados experimentais apontam, no contexto deste teste, para um valor de cerca de 3 ms para a soma de $T_{getSNMP}$, $T_{respSNMP}$ and T_{AgSNMP}.

$$T_{EC} = N \cdot \left(T_{getSNMP} + \frac{S_{get\_pdu}}{BW} + T_{AgSNMP} + \frac{S_{resp\_pdu}}{BW} + T_{respSNMP} \right) \quad (1)$$

$$T_{MW} = \frac{S_{arranque}}{BW} + N \cdot \left(T_{getSNMP} + T_{AgSNMP} + T_{respSNMP} \right) + N \cdot \left(\frac{f_{compressão} \cdot S_{resp\_pdu}}{BW} \right) \quad (2)$$

$$\frac{T_{EC}}{T_{MW}} = \frac{K \cdot BW + S_{get\_pdu} + S_{resp\_pdu}}{K \cdot BW + \frac{S_{Arranque}}{N} + f_{compressão} \cdot S_{resp\_pdu}}, \text{ sendo } K = T_{getSNMP} + T_{AgSNMP} + T_{respSNMP}, \quad (3)$$

BW representa a largura de banda disponível, S_{get_pdu} e S_{resp_pdu} indicam o tamanho das mensagens SNMP (no nosso caso específico 122 e 128 octetos, respectivamente) e $S_{arranque}$ corresponde ao tamanho da mensagem inicial enviada pelo nó central para desencadear um novo ciclo de consulta, no modelo MW (64 octetos). A relação entre o tamanho das respostas SNMP e o tamanho dos dados enviados para o ponto central no modelo MW é indicada por $f_{compressão}$. As nossas medições mostraram que este factor varia consoante a quantidade total de dados envolvidos: 67% para 5 transacções SNMP, 44% para 10 transacções, 29% para 20, 21% para 40 e 18% para 80 e 100 transacções SNMP. Ainda que a compressão aplicada aos dados apontasse para uma taxa estável de 20%, os *overheads* de comunicação justificam esta variação. O segundo gráfico da Figura 2, que representa os valores analíticos obtidos

aplicando a Expressão 3*, mostra que apesar de alguma irregularidade nas medições experimentais existe correspondência entre as medições experimentais e os valores analíticos.

A primeira conclusão a destacar destes resultados é que para um dado valor de latência de rede existe um limite fixo para os ganhos relativos obtidos enviando um agente móvel para o nó gerido, sendo o aumento da compressão destes dados a única forma de alargar esse limite. Outra conclusão relevante é que os ganhos de desempenho obtidos enviando o agente móvel para o nó gerido são relativamente reduzidos em ambientes de rede local: se bem que com 512 Kbps os ganhos sejam de cerca de 289%, com 10 Mbps eles ficam já abaixo de 10%, um valor marginal que provavelmente não justificará os custos acrescidos do modelo MW.

3.2 O Efeito da Distribuição

Quando existe mais de um nó gerido o ponto central torna-se num potencial ponto de estrangulamento. Mesmo considerando transacções SNMP assíncronas – no sentido em que é possível enviar pedidos aos nós B e C antes de receber a resposta do nó A – os recursos computacionais do ponto central tendem a ficar sobrecarregados, conduzindo à degradação dos tempos de resposta. Nesta situação o modelo MW permite aumentar o desempenho do sistema, uma vez que a carga computacional é separada por diversos agentes móveis que trabalham em paralelo em pontos distintos da rede. Poderá continuar a ocorrer congestão no envio final dos dados para o ponto central, mas devido ao prévio processamento e compressão desses dados o efeito dessa congestão será menos determinante.

A Figura 3 apresenta algumas das medições de desempenho para os modelos EC e MW, numa rede comutada a 10 Mbps. Os resultados mostram que o tempo de resposta do modelo EC é quase proporcional ao número de nós geridos, indicando que os recursos computacionais do ponto central constituem um ponto de estrangulamento. O modelo MW, por seu lado, apresenta uma menor degradação com o aumento de dimensão da rede. A análise de um conjunto mais alargado de resultados mostrou que por cada nó adicional a degradação do desempenho, relativamente ao tempo de resposta obtido com um único nó gerido, é de apenas 2,65% (i.e. o tempo de resposta com 10 nós geridos é aproximadamente 26% mais elevado que o tempo de resposta com 1 nó gerido, e com 20 nós geridos será aproximadamente 52% mais elevado). Esta degradação justifica-se pela competição entre os diversos nós geridos na transferência final dos dados processados para o ponto central.

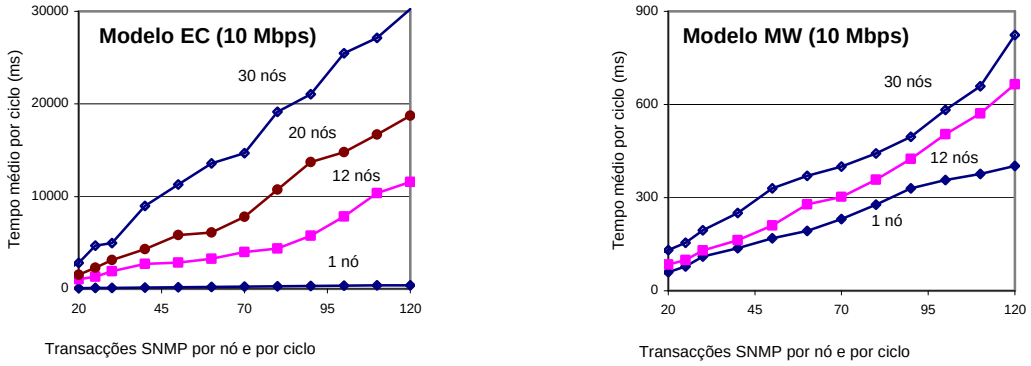


Figura 3: Desempenho dos Modelos EC e MW (vários nós geridos)

Estas medições não são surpreendentes, ainda que possam porventura ser impressionantes. Elas mostram que em ambientes de rede local é a *distribuição* que mais contribui para o desempenho, e não a *localidade* das transacções SNMP. Esta é uma conclusão importante porque a localidade é muitas vezes difícil de obter, uma vez que colocar um agente móvel ou

* Note-se que na Expressão 3 se usa a largura de banda como forma de simplificar a determinação da latência da rede. Na nossa experiência existia um *router* entre o ponto central e o nó gerido, sendo a largura de banda controlada nos dois interfaces do *router*. Isto duplicou a latência, afectando em especial o modelo EC, no qual são trocadas pela rede sucessivas sequências pedido/resposta. Por esta razão os valores da Figura 2 pressupoem que, no caso do modelo EC, a latência será aproximadamente *2S/BW*, e não *S/BW*.

qualquer outro tipo de código móvel em todos os dispositivos da rede pode ser impraticável, demasiado oneroso ou impedido por razões de segurança e portabilidade. A distribuição, por outro lado, pode ser obtida com menores custos usando um conjunto de nós estrategicamente seleccionados, cada um deles gerindo uma pequena parte da rede, como sucede com o modelo ED. Neste contexto, torna-se possível ajustar os níveis de distribuição – o número de gestores delegados – de acordo com os recursos disponíveis e o desempenho pretendido.

A Figura 4 apresenta algumas medições de desempenho para o modelo ED. O primeiro gráfico mostra como o desempenho pode ser ajustado – entre os limites mínimo e máximo estabelecidos pelos modelos EC e MW – seleccionando um número apropriado de domínios de delegação. O segundo gráfico mostra como é possível manter bons níveis de desempenho à medida que aumenta a dimensão da rede, criando para o efeito mais domínios de delegação em vez de aumentar a dimensão de cada domínio. O tempo médio de resposta do modelo ED com 120 nós geridos (repartidos por 12 domínios) é inferior ao dobro do tempo medido para o modelo EC com apenas 10 nós geridos.

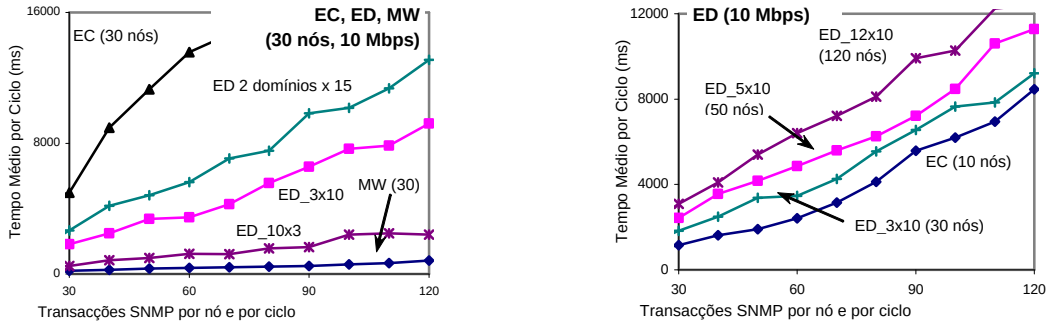


Figura 4: Desempenho do Modelo ED

4. Desempenho de Modelos Migratórios (MG, MD)

Enquanto na Secção 3 foi discutido o efeito da localidade e da localidade no desempenho da gestão distribuída, nesta Secção é introduzido um terceiro factor: a mobilidade. Do ponto de vista do desempenho a mobilidade corresponde a uma forma alternativa de obter localidade. Enquanto o modelo MW se baseia em mobilidade restrita e forte distribuição – um AM estacionado em cada nó gerido – os modelos migratórios usam a mobilidade como forma alternativa de conduzir transacções SNMP locais com menor grau de distribuição (modelo MD) ou mesmo sem nenhuma distribuição do processo de gestão (modelo MG). Contudo, uma vez que a migração dos AM é um processo relativamente lento, os modelos migratórios apresentarão à partida custos de desempenho mais elevados, mesmo considerando técnicas especificas de aceleração da migração [26].

Os resultados experimentais (Figura 5) confirmam esta noção, mostrando que o desempenho dos modelos migratórios não é competitivo, mesmo em comparação com o modelo EC. Como é óbvio uma forte paralelização do processo melhora o seu desempenho (vejam-se por exemplo as medições obtidas para o modelo MD com 10 domínios de 2 nós cada) mas, ainda assim, os modelos estáticos com grau equivalente de distribuição serão em geral mais rápidos.

5. Medições de Tráfego

Ao efectuar transações SNMP localmente, alguns modelos de distribuição (MG, MD e MW) podem processar e agregar os dados antes de os transmitir pela rede, reduzindo o seu volume directamente na fonte. Mecanismos como a compressão semântica [20] ou a filtragem de tabelas [21] permitirão assim diminuir significativamente o tráfego de rede: taxas de compressão de 1 para 5 são provavelmente possíveis na maior parte das aplicações, e taxas bastante mais elevadas poderão ser obtidas em algumas aplicações de monitorização.

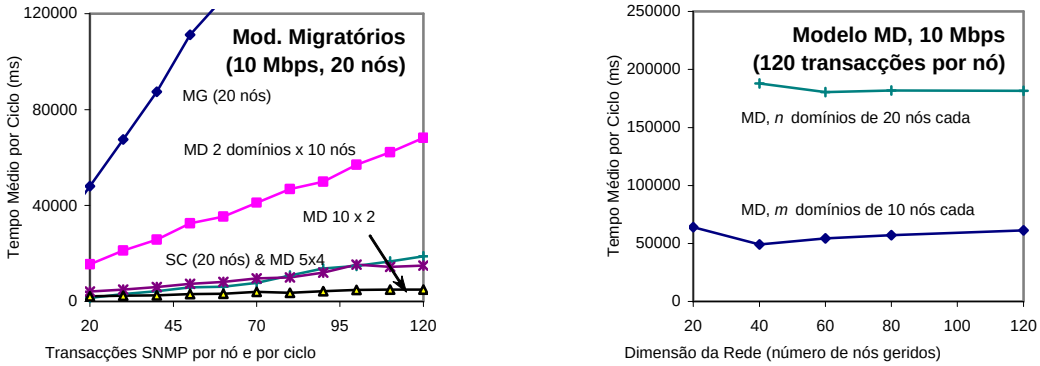
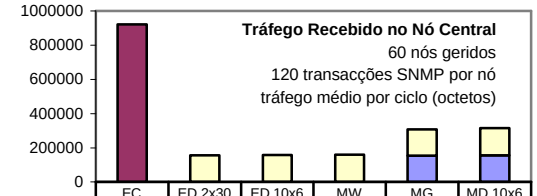


Figura 5: Desempenho e Escalabilidade dos Modelos Migratórios (MG, MD)

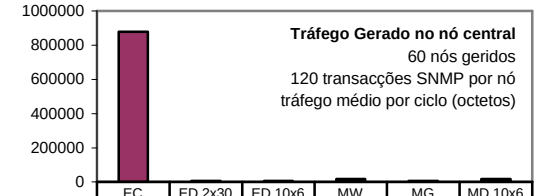
Noutros modelos, tais como o ED, a compressão de dados ocorre apenas num nível intermédio: o gestor delegado. Esta situação é pouco interessante se os custos do tráfego trocado entre o gestor delegado e o nó gerido forem equivalentes aos custos do tráfego trocado entre o gestor delegado e o ponto central, uma vez que o tráfego global corresponderá à soma do tráfego associado às transacções SNMP com a transmissão final dos dados para o ponto central. No entanto, quando os domínios de delegação não cruzam fronteiras de redes locais os custos do tráfego interno são tipicamente menos relevantes. Nestes ambientes, em que os custos de comunicação são determinados essencialmente pelas ligações entre o domínio de delegação e o ponto central, os benefícios da compressão de dados do modelo ED poderão ser similares aos benefícios dos modelos baseados em localidade, tais como o MW.

A Figura 6 mostra as medições médias de tráfego recebido e gerado no ponto central para um ciclo "regular" (i.e. os custos de arranque do primeiro ciclo não são incluídos), considerando três categorias distintas: transacções SNMP; transferência de informação de gestão previamente processada; e mensagens associadas ao controlo da infra-estrutura de suporte aos agentes móveis. Nos resultados apresentados destacam-se três aspectos: o efeito da compressão de dados; o peso do controlo da infra-estrutura nos modelos migratórios; e a ineficiência do SNMP em termos de tráfego gerado pelo ponto central.

Conforme foi já mencionado, a nossa aplicação de testes aplica técnicas de compressão que reduzem a informação de gestão para 20% do seu tamanho original. Contudo, a redução efectiva de tráfego na transferência de informação de gestão – em comparação com transações SNMP directas – é também influenciada pelos *overheads* protocolares. Os resultados apresentados na Figura 6 consideram 120 transações SNMP (i.e. 600 objectos) por nó e por ciclo e, por conseguinte, mostram uma taxa de compressão bastante favorável: entre os 16,7% do tamanho original para o modelo MG (onde são agregados os dados de todos os nós geridos antes do envio) e os 17,5% para o modelo MW (onde a agregação ocorre ao nível da cada nó). No entanto, tal como seria de esperar, as taxas de compressão efectivamente medidas para menores volumes de informação são inferiores, tal como se mostra na Tabela 1.



	EC	ED 2x30	ED 10x6	MW	MG	MD 10x6
Inf.Gestão	0	154256	158567	161027	154094	161110
SNMP	920811	0	0	0	0	0
Controlo	0	14	21	64	153437	154230



	EC	ED 2x30	ED 10x6	MW	MG	MD 10x6
Inf.Gestão	0	4729	5767	15353	4608	7130
SNMP	877612	0	0	0	0	0
Controlo	0	522	619	1061	1342	9896

Figura 6: Tráfego de Rede (medido junto ao nó central)

Transacções SNMP por ciclo	*5*	*10*	*20*	*40*	*80*	*120*
Taxa de Compressão (MW / EC)	67%	44%	29%	21%	18%	18%
Taxa de Compressão (ED 4x5 / EC)	33%	25%	22%	19%	18%	17%
Taxa de Compressão (ED 2x10 / EC)	24%	22%	19%	18%	17%	17%

Tabela 1: Taxas de Compressão Efectivas (20 nós geridos, taxa nominal de 20%)

No caso dos modelos migratórios os custos associados ao controlo da infra-estrutura de agentes móveis foram consideráveis, uma vez que o nó central recebe notificações detalhadas sempre que ocorre a migração de um AM. Contudo, esta circunstância depende na forma específica como a plataforma JAMES foi usada nestes testes, e por conseguinte outras implementações poderão facilmente reduzir este tráfego em troca de mecanismos de monitorização menos exigentes.

As diferenças entre o modelo EC e os restantes modelos, em termos de tráfego gerado pelo ponto central, são explicadas pela ineficiência do protocolo SNMP, que torna necessário enviar sucessivos pedidos de consulta (comandos *get* ou *get-next*), aumentando desnecessariamente o volume de tráfego (pelo menos nas situações em que não seja possível usar o comando *get-bulk*). Com gestão distribuída torna-se possível reduzir o volume tráfego gerado ou, pelo menos, desvia-lo do nó central para os gestores delegados. Deve no entanto ser mencionado que as medidas de tráfego apresentadas neste artigo correspondem à situação mais favorável, em que os gestores remotos já sabem, à partida, que informação recolher em cada ciclo – a aplicação central limita-se a desencadear o inicio do ciclo de consultas e a confirmar a recepção dos dados correspondentes. Outras aplicações distribuídas poderão implicar maiores fluxos de informação do nó central para os gestores remotos.

6. Sensibilidade às Condições de Rede

As medições de desempenho apresentadas nas Secções 3 e 4 foram obtidas usando um ambiente típico de rede local (rede comutada a 10 Mbps) e, tal como foi já referido, eles dependem essencialmente do grau de distribuição do processo de gestão. Será contudo de prever que a localidade assuma gradualmente uma maior relevância para o desempenho do sistema à medida que diminui a largura de banda disponível. Para analisar essa situação, foram também efectuadas medições de desempenho com maiores restrições de largura de banda na ligação entre os nós geridos e a aplicação central.

No caso dos modelos migratórios (MG, MD) a redução da largura da banda teve um efeito praticamente imperceptível no desempenho do sistema, provavelmente porque a degradação do canal entre o ponto central e os nós geridos não interfere com a migração dos AM e, por conseguinte, apenas a transmissão final dos dados sofre algum impacto.

Já nos modelos de mobilidade restrita as restrições de largura de banda são mais penalizadoras. A Figura 7 apresenta o tempo médio por ciclo para os modelos MW e ED, considerando 20 nós geridos. Enquanto a rede mantém débitos relativamente elevados a degradação do desempenho é perceptível mas relativamente reduzida, uma vez que o atraso suplementar verificado no envio final dos dados para o ponto central é ainda relativamente reduzido. Considerando a largura de banda nominal e os volumes de dados transmitidos, a degradação esperada com a passagem de 2 Mbps para 1 Mbps, por exemplo, é de cerca de 180 ms (valores analíticos considerando 100 transacções SNMP). Isto corresponderia a uma degradação de cerca de 22% para o modelo MW e 6% para o modelo ED com 4 domínios.

Contudo, o peso relativo desta degradação aumenta gradualmente até ao ponto em que se torna o principal factor no desempenho global do sistema. Os tempos médios medidos para o modelo MW, por exemplo, são 3 vezes piores com 256 Kbps que com 2 Mbps. A dado ponto o tempo de processamento torna-se praticamente irrelevante e a capacidade dos canais de comunicação assume-se como principal ponto de estrangulamento do sistema. A Figura 8

ilustra esta perspectiva comparando as medições de desempenho do modelo EC com os valores máximos teóricos calculados apenas com base na largura de banda nominal e no tráfego de rede (i.e. assumindo atrasos de processamento nulos). Com os modelos MW e ED foram registados resultados similares.

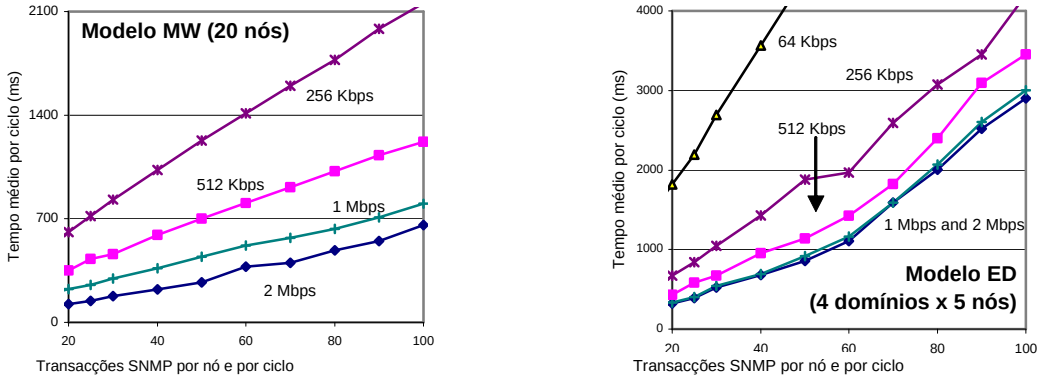


Figura 7: Efeito da Restrição da Largura de Banda nos Modelos de Mobilidade Restrita

Neste cenário de escassez de largura de banda o desempenho de cada modelo torna-se dependente da sua capacidade de comprimir a informação de gestão. Neste sentido, a *localidade* assume um papel importante por possibilitar a agregação e compressão de dados directamente na fonte, mas a *proximidade* – na forma como é usada no modelo ED – pode produzir resultados similares, desde que os custos de comunicação intra-domínio sejam substancialmente inferiores aos custos de transmissão entre os gestores delegados e o ponto central. De qualquer modo, o importante é comprimir os dados antes das ligações críticas: se os dados forem comprimidos para metade do tamanho original os tempos de resposta poderão também ser reduzidos para próximo de metade dos valores originais.

Importa também clarificar uma noção generalizada de que – relativamente a modelos centralizados ou menos distribuídos – as soluções distribuídas se comportam tanto melhor quanto menos favoráveis forem as condições de rede. Soluções fortemente distribuídas são concerteza interessantes nestes ambientes, dado o seu maior potencial na redução do tráfego de rede. Contudo, é importante ter presente que o seu desempenho relativo – comparado com soluções mais centralizadas – pode mesmo ser superior em redes mais favoráveis: aplicações fortemente distribuídas que, por alguma razão, não consigam comprimir significativamente os dados trocados pela rede terão menos vantagens relativas à medida que as condições da rede piorem. Esta observação é óbvia mas frequentemente esquecida, uma vez que a maior parte dos estudos experimentais usa redes de reduzida dimensão onde os benefícios da distribuição não são tão evidentes como os benefícios da compressão de dados.

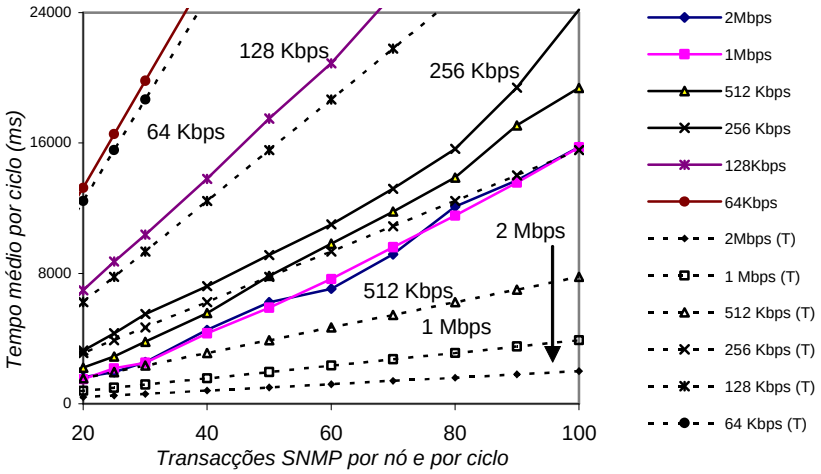


Figura 8: Desempenho Medido e Máximo Desempenho Teórico (EC, 20 nós geridos)

7. Custos de Arranque e Instalação

Os estudos apresentados nas Secções anteriores focam-se no funcionamento do sistema em "estado estacionário", quando os agentes delegados estão já distribuídos pela rede e quando os mecanismos de *caching* optimizam a migração dos AM. Foi assim assumido que para a maioria das aplicações práticas o sistema seria usado durante um período de tempo suficiente para compensar os eventuais custos de arranque de cada modelo. No entanto, não deixa de ser importante ter uma noção de quais são efectivamente esses custos em cada situação.

Os custos de arranque diferem de modelo para modelo. Os modelos migratórios (MG, MD) tendem a ser mais lentos no primeiro ciclo de recolha de informação, quando os mecanismos de optimização de migração não podem ainda tirar partido de *caches* de código. No entanto, nos testes efectuados a diferença entre o desempenho do primeiro ciclo e os ciclos seguintes raramente excedeu os 10%. Já nos modelos de mobilidade restrita se registou uma diferença muito maior entre o primeiro ciclo – no qual os gestores remotos são instalados – e os ciclos seguintes.

Os dois gráficos da Figura 9 apresentam os tempos médios de resposta acumulados para diversos modelos, considerando um ambiente de 20 nós geridos e, respectivamente, 30 e 120 transacções SNMP por ciclo e por nó (ou seja, 150 e 600 objectos SNMP por nó e por ciclo). No primeiro caso (150 objectos) a diferença absoluta entre o desempenho de cada modelo é inferior, e por conseguinte o ponto de compensação para o modelo mais rápido (MW) ocorre apenas por volta do 25º ciclo. No segundo caso os maiores volumes de dados resultam em maiores diferenças entre os diversos modelos, e por conseguinte os custos de arranque, que se mantém constantes, são compensados mais cedo (por volta 11º ciclo). Note-se, no entanto, que os modelos mais lentos, tais como o EC, são rapidamente ultrapassados, mesmo considerando um número reduzido de ciclos. Do ponto de vista do tráfego de rede as medições efectuadas produziram resultados similares.

Um outro aspecto relacionado com os custos de arranque e instalação prende-se com os requisitos impostos aos nós geridos. Os modelos descentralizados discutidos neste artigo necessitam de ser suportados por uma infra-estrutura de agentes móveis espalhada pela rede. Em comparação com o modelo clássico centralizado – que apenas exige aos nós geridos o suporte da norma SNMP – esta infra-estrutura de agentes móveis exige mais recursos computacionais, é mais difícil de instalar e requer manutenção explicita. Em algumas situações, a instalação e manutenção duma infra-estrutura deste tipo é simplesmente impossível (por exemplo no caso de sistemas embebidos ou proprietários), indesejável ou demasiado onerosa. Nesta perspectiva o modelo EC encontra-se em vantagem, pois não necessita de qualquer tecnologia adicional para além do "omnipresente" SNMP. Os modelos MW, MG e MD pressupõem que todos os nós geridos suportam tecnologia de agentes móveis. O modelo ED, por seu lado, representa um compromisso menos exigente, pois esse suporte é necessário apenas por parte dos nós onde ficam instalados os gestores delegados.

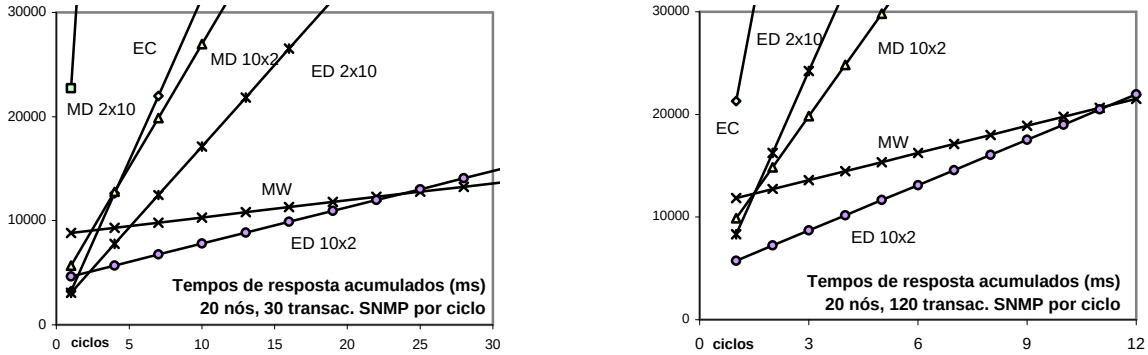


Figura 9: Tempos de Resposta Acumulados (pontos de compensação dos custos de arranque)

8. Conclusões

Nos últimos anos diversos estudos abordaram a eficiência dos sistemas de gestão distribuída baseados na tecnologia de agentes móveis. Estes estudos demonstraram as suas potenciais vantagens em vertentes como a redução do tráfego de rede, o desempenho e a escalabilidade. No entanto, apesar desta relativa abundância de estudos, consideramos que o comportamento de sistemas de agentes móveis em cenários típicos de gestão não é ainda plenamente compreendido. Esta convicção motivou a condução de um estudo experimental que reproduz uma gama alargada de cenários de gestão. Em vez de focar um aspecto específico, neste artigo tentou-se deliberadamente apresentar uma visão global sobre um conjunto relativamente alargado de vertentes e sobre a forma como estas vertentes interagem entre si. Esta visão alargada não conduz a resultados substancialmente diferentes dos de estudos anteriores mas fornece, em alternativa, uma perspectiva bastante diferente desses resultados.

Uma das principais lições deste estudo foi a percepção de que, num largo conjunto de situações (incluindo os ambientes de rede local) os ganhos de desempenho são determinados essencialmente pela *distribuição* do processo de gestão, e não pela *localidade* (condução de transacções SNMP locais), cujos ganhos são marginais. Existe em geral a percepção, correcta, de que a localidade e a distribuição determinam, em conjunto, o desempenho dos sistemas de agentes móveis. Contudo, não esperávamos um desequilíbrio tão grande entre estes dois factores numa simples rede local a 10 Mbps. Esta questão é relevante porque existem formas alternativas de obter distribuição, tais como o modelo ED – cujo nível de desempenho pode ser ajustado em função dos custos de espalhar agentes delegados pela rede.

Naturalmente, com condições de rede mais severas o peso relativo da *distribuição* diminui, atingindo eventualmente um ponto a partir do qual o desempenho é determinado quase exclusivamente pelo tempo gasto a transferir os dados pela rede (e não pelo tempo gasto a processar esses dados). Neste caso a *localidade* tem um papel relevante na medida em que permite comprimir os dados directamente na fonte mas, quando os canais de comunicação intra-domínio são muito melhores que os canais de comunicação até ao ponto central, a *localidade* pode ser substituída pela *proximidade* (no sentido em que é usada pelo modelo ED) sem perdas de desempenho. Nestas situações *localidade* e *proximidade* são apenas meios alternativos de reduzir o volume da informação de gestão antes do troço de rede mais crítico.

Tradicionalmente a *localidade* é também apontada como solução para os atrasos resultantes da latência de rede (provocados por sucessivas consultas SNMP). Contudo, se a aplicação de gestão consulta diversos nós geridos em simultâneo, como habitualmente sucede, o efeito da latência de rede torna-se menos relevante que o débito da rede.

Outra lição importante deste estudo foi a constatação de que, do ponto da vista do desempenho, a compressão de dados é quase irrelevante com boas condições de rede. De facto, algumas experiências não discutidas neste artigo mostraram-nos que com 10 Mbps de largura de banda o uso de técnicas de compressão mais agressivas não trazem melhorias sensíveis de desempenho, mesmo quando não aumentam os requisitos de processamento. Contudo, com a gradual escassez de largura de banda a compressão de dados passa a ser determinante para o desempenho do sistema, e cada modelo de distribuição pode tornar-se tão rápido quanto a taxa de compressão que consegue atingir.

Por ultimo, importa salientar que os modelos migratórios (MG, MD) não são superiores aos modelos baseados em mobilidade restrita (MW, ED), quer em termos de desempenho quer em termos de ocupação da rede. No entanto, uma vez que os modelos migratórios são apenas uma de várias formas possíveis de aplicar agentes móveis na gestão distribuída, os maus resultados obtidos por esses modelos não devem ser generalizados ao paradigma de agentes móveis: o nosso estudo mostra que os agentes móveis, quando associados a modelos de distribuição apropriados, permitem construir aplicações com excelentes níveis de desempenho. Além

disso, ainda que os modelos de mobilidade restrita possam também ser construídos sobre outras tecnologias de programação distribuída – potencialmente capazes de fornecer níveis de desempenho idênticos – é nossa convicção que o uso de agentes móveis se justifica também por permitir melhores metáforas de programação e maior flexibilidade.

Referências

[1] P. Simões, J. Rodrigues, L. Silva, F. Boavida, "Distributed Retrieval of Management Information: Is it About Mobility, Locality or Distribution?", Proceedings of the 2002 IEEE/IFIP Network Operations and Management Symposium (NOMS'2002), IEEE Press, Abril de 2002.

[2] J.-F. Martin-Flatin, S. Znaty, "Two Taxonomies of Distributed Network and Systems Management Paradigms", Technical Report DSC/2000/032, École Polytechnique Fédérale de Lausanne, Suiça, 2000.

[3] Y. Yemini, G. Goldszmidt, S. Yemini, "Network Management by Delegation", Proceedings of IFIP 2nd International Symposium on Integrated Network Management (ISINM'91), Washington, 1991.

[4] M. Siegl, G. Trausmuth, "Hierarchical Network Management: A Concept and its Prototype in SNMPv2", Computer Networks and ISDN Systems, Vol. 28, No. 4, pp. 441-452, 1996.

[5] J. Schönwälder, J. Quittek, C. Kappler, "Building Distributed Management Applications Using the IETF Script MIB", IEEE Journal on Selected Areas in Communications, Vol. 18, Nº. 5, pp. 702-714, IEEE Comm. Society, Maio de 2000.

[6] A. Schade, A. P. Trommler, "CORBA-based Model for Distributed Application Management", Proceedings of the 7th IFIP/IEEE Workshop on Distributed Systems: Operations & Management (DSOM'96), 1996.

[7] A. Hayzelden, J. Bigham (Eds.), "Software Agents for Future Communication Systems", Springer-Verlag, 1999.

[8] N. Anerousis, "Scalable Management Services Using Java and the World Wide Web", Proc. of the 9th IFIP/IEEE Int. Workshop on Distributed Systems: Operations and Management (DSOM'98), Delaware, 1998.

[9] G. Aschemann, S. Domnitcheva, P. Hasselmeyer, R. Kehr, A. Zeidler, "A Framework for the Integration of Legacy Devices into a Jini Management Federation", Proc. of the Tenth IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM'99), Springer-Verlag, 1999.

[10] A. Bieszcad, B. Pagurek, T. White, "Mobile Agents for Network Management", IEEE Communications Surveys, 4Q, 1998.

[11] M. Baldi, G. Picco, "Evaluating the Tradeoffs of Mobile Code Design Paradigms in Network Management Applications", Proceedings of ICSE'98 – 20th International Conference on Software Engineering, 1998.

[12] M. Rubinstein, O. Duarte, "Evaluating Tradeoffs of Mobile Agents in Network Management", Networking and Information Systems Journal, Vol. 2, No. 2, 1999, Hermes-Science Publications.

[13] A. Liotta, G. Knight, G. Pavlou, "On the Performance of Decentralised Monitoring using Mobile Agents", Proc. of the Tenth IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM'99), Springer-Verlag, 1999.

[14] R. Pinheiro, A. Poylisher, H. Caldwell, "Mobile Agents for Aggregation of Network Management Data", Proceedings of ASA/MA'99 – First International Symposium on Agent Systems and Applications and First International Symposium on Mobile Agents, pp. 130-141, Palm Springs, 1999.

[15] M. Zapf, K. Herrmann und K. Geihs, "Decentralized SNMP Management with Mobile Agents", Proceedings of the IM'99, Boston, 1999.

[16] A. Puliafito, O. Tomarchio, "Using Mobile Agents to Implement Flexible Network Management Strategies", Computer Communication Journal, 23(8), Abril de 2000.

[17] A. Sahai, C. Morin, "Enabling a Mobile Network manager (MNM) Through Mobile Agents", Proceedings of Mobile Agents'98 (MA'98), Estugarda, 1998.

[18] C. Bohoris, A. Liotta, G. Pavlou, "Evaluation of Constrained Mobility for Programmability in Network Management", Proceedings of the 11th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM'2000), Springer-Verlag LNCS 1960, 2000.

[19] S. Lipperts, "How to Efficiently Deploy Mobile Agents for an Integrated Management", Proc. of the 3rd IFIP/GI Int. Conference on Trends towards a Universal Service Market (USM'2000), Munique, 2000.

[20] D. Gavalas, D. Greenwood, M. Ghanbari, M. O'Mahony, "An Infrastructure for Distributed and Dynamic Network Management based on Mobile Agent Technology", Proc. of the IEEE Int. Conf. on Communications (ICC'99), 1999.

[21] D. Gavalas, D. Greenwood, M. Ghanbari, M. O'Mahony, "Enabling Mobile Agent Technology for Intelligent Bulk Management Data Filtering", Proceedings of the 2000 IEEE/IFIP Network Operations and Management Symposium (NOMS'2000), Honolulu, Hawaii, 2000.

[22] D. Gavalas, D. Greenwood, M. Ghanbari, M. O'Mahony, "Implementing a Highly Scalable and Adaptive Agent-Based Management Framework", Proc. of the IEEE Global Comm. Conference (Globecom'2000), São Francisco, 2000.

[23] L. Silva, P. Simões, G. Soares, P. Martins, V. Batista, C. Renato, L. Almeida, N. Stohr, "JAMES: A Platform of Mobile Agents for the Management of Telecommunication Networks", Proc. of IATA'99, Springer-Verlag, 1999

[24] P. Simões, L. Silva, F. Boavida, "Integrating SNMP into a Mobile Agent Infrastructure", Proc. of the 10th IFIP/IEEE Int. Workshop on Distributed Systems: Operations and Management (DSOM'99), Springer-Verlag, 1999.

[25] L. Rizzo, "Dummynet: a simple approach to the evaluation of network protocols", ACM Computer Communication Review, Vol. 27, No. 1, pp. 31-41, Janeiro de 1997.

[26] L. Silva, V. Batista, P. Martins, G. Soares, "Using Mobile Agents for Parallel Processing", Proceedings of DOA'99 (Int. Symposium on Distributed Objects and Applications), IEEE Computer Press, 1999.