1
Introdução

O termo Qualidade de Serviço (QoS) tem sido muito referenciado nos últimos anos, devido, principalmente, à crescente demanda por aplicações distribuídas de alto desempenho, que exigem certos requisitos do sistema, como retardo máximo, variação estatística máxima do retardo, taxa mínima de transmissão, entre outros. Oferecer um serviço com QoS significa permitir que tais parâmetros sejam garantidos às aplicações, de acordo com os valores por elas fornecidos, durante seu período de execução.

A provisão de qualidade de serviço em sistemas caracterizados por fortes exigências de processamento e comunicação impõe a necessidade de atendimento aos requisitos de maneira fim-a-fim, ou seja, cada subsistema componente do serviço deve ter seus recursos gerenciados no intuito de garantir aos usuários o nível de qualidade por cada um deles solicitado (Gomes, 1999). Numa aplicação multimídia distribuída, por exemplo, não só o sistema de comunicação em rede deve oferecer os mecanismos para a provisão de QoS, mas também o sistema operacional das estações envolvidas e, recursivamente, os seus subsistemas relevantes. Nas estações finais, recursos controlados pelo sistema operacional, como CPU, memória e buffers de comunicação, devem ser gerenciados de forma a assegurar que a coexistência de várias aplicações não viole as necessidades individuais de QoS de cada uma delas.

Para provisão de QoS, é interessante que os sistemas envolvidos, em particular os sistemas operacionais, sejam flexíveis, para que novos serviços possam ser configurados visando sua utilização por futuras aplicações. A especificação de um novo serviço pode envolver a escolha de algoritmos de escalonamento, admissão e classificação, por exemplo, no que se refere diretamente ao uso dos recursos de processamento e comunicação. Outros parâmetros de configuração podem ser citados, como as tarefas que irão compor a pilha de protocolos de comunicação ou a descrição do estado inicial do sistema

para a provisão de QoS (conjunto de algoritmos citados e particionamento inicial dos recursos para cada tipo de aplicação). Em outras palavras, é desejável que os sistemas possam ser adaptáveis, para o oferecimento de novas modalidades de serviço, a partir da modificação de seu estado interno em tempo de operação.

O processamento por compartilhamento do tempo e a estrutura monolítica em que se baseiam os sistemas operacionais de uso geral criam, contudo, várias barreiras para adaptabilidade[1] e provisão de QoS. Os escalonadores de processos desses sistemas recorrem apenas ao expediente do uso de prioridades para privilegiar a execução de algumas aplicações em detrimento de outras. Seus subsistemas de rede são guiados por interrupções, o que pode causar anomalias de escalonamento de processos. Normalmente, as filas de transmissão de pacotes são compartilhadas, não havendo meios para a classificação ou priorização dos pacotes. Por último, são raros os mecanismos para a reserva de recursos e para a introdução de partes adaptáveis no *kernel*[2] em tempo de execução.

Visto que os sistemas operacionais de uso geral possuem pouco ou nenhum suporte para a adaptabilidade e provisão de QoS, muitas foram as pesquisas desenvolvidas nessas áreas. Alguns trabalhos integram projetos para a construção de novos sistemas, estruturados desde suas concepções para desempenharem tais funções. Outros trabalhos propõem a extensão de sistemas operacionais já existentes, sejam eles de uso geral ou não. Examinando-se vários desses estudos, percebe-se que os mecanismos propostos possuem semelhanças funcionais entre si, o que evidencia a possibilidade de serem descritos de forma genérica. O desenvolvimento de uma arquitetura genérica para provisão de QoS é de grande utilidade, uma vez que pode clarificar os conceitos envolvidos, facilitar a reutilização das funcionalidades comuns e definir uma organização interna que seja equivalente nos diferentes sistemas.

---

[1] Nota-se que o termo adaptabilidade, no contexto de qualidade de serviço, é utilizado em muitos trabalhos para descrever a capacidade de reação dos subsistemas componentes do serviço mediante variações apresentadas na carga sobre os recursos envolvidos. Esse conceito será introduzido no Capítulo 3, com a denominação de sintonização da QoS, evitando qualquer ambigüidade sobre o termo "adaptabilidade" neste trabalho.

[2] Neste trabalho, será usado o substantivo kernel da língua inglesa, para denominar o núcleo do sistema operacional, ou seja, o conjunto de rotinas que integram o controle central da estação.

1.1 Objetivos da Dissertação

O objetivo principal deste trabalho é a descrição de uma arquitetura genérica adaptável para a provisão de QoS nos subsistemas de rede e de escalonamento de processos de sistemas operacionais, baseada nos *frameworks para provisão de QoS em ambientes genéricos de processamento e comunicação,* apresentados em (Gomes, 1999).

O trabalho referenciado acima identifica as funções recorrentes de provisão de QoS nos vários ambientes envolvidos (e.g. redes de comunicação, sistemas distribuídos e sistemas operacionais) e como essas funções participam da orquestração de recursos para o fornecimento de serviços com QoS verdadeiramente fim-a-fim. Os componentes foram estruturados sob a forma de frameworks no intuito de facilitar a identificação dos pontos de flexibilização (*hot-spots*), que devem ser preenchidos para descrever a funcionalidade de um ambiente específico.

Particularmente, o presente trabalho mostra como alguns desses pontos de flexibilização podem ser completados para acomodar várias técnicas de provisão de QoS em sistemas operacionais. Por outro lado, outros pontos de flexibilização são deixados em aberto para que a arquitetura atenda aos requisitos de generalidade e de adaptabilidade a novos serviços, possibilitando a configuração de certas funções já citadas anteriormente.

Outro objetivo deste trabalho é mostrar que o sistema Linux (Rusling, 1996), apesar de caracterizado pelas desvantagens comuns aos sistemas operacionais de uso geral, pode ser ligeiramente modificado para que alguns mecanismos de provisão de QoS possam ser oferecidos. Dessa forma, a arquitetura proposta pode ser aplicada em um cenário de uso sobre este sistema de código aberto.

O enfoque do presente trabalho, definido sobre os subsistemas de comunicação e processamento de sistemas operacionais, foi motivado pela importância da provisão de QoS para aplicações multimídia distribuídas, caracterizadas pelo uso em massa dos recursos desses dois subsistemas. Nota-se,

porém, que recursos como memória principal, memória virtual (sistema de paginação), discos (memória secundária), interfaces de vídeo e de som são, também, muito importantes para esse tipo de aplicação, o que torna necessária uma orquestração também entre esses recursos. Por ser um assunto extremamente amplo, optou-se pela restrição citada acima, fundamentada no fato de que diversos trabalhos relacionados apresentaram grande aumento na eficiência do sistema como um todo, propondo soluções apenas para o escalonamento de processos e o processamento da pilha de protocolos.

Reforçando este empirismo, é observado que existe uma ordem natural para o desenvolvimento de mecanismos de provisão de QoS em sistemas operacionais. Por exemplo, menor será o ganho na implementação de um subsistema de paginação de memória com garantias de QoS se, antecipadamente, não forem tratadas as questões acerca da própria execução dos processos. O gerenciamento eficiente da preempção das aplicações e do processamento da pilha de protocolos pode, adicionalmente, levar a uma redução do número de mudanças de contexto e, conseqüentemente, do número de páginas a serem buscadas na memória virtual.

1.2 Estrutura da dissertação

Esta dissertação encontra-se estruturada como se segue. No Capítulo 2, são descritas as características dos subsistemas de rede e de escalonamento de processos que são relevantes no estudo da provisão de qualidade de serviço em ambientes multimídia distribuídos. Primeiramente, é feito um rápido estudo sobre os sistemas operacionais de uso geral, para que possam ser apontadas as particularidades que representam empecilhos para a provisão de QoS. Em seguida, são apresentados alguns trabalhos relacionados a esses pontos críticos, além de arquiteturas alternativas para a construção de sistemas operacionais com suporte a qualidade de serviço. Essas informações são importantes para que sejam identificados os pontos de flexibilização dos frameworks propostos e para que seja construída uma modelagem genérica do funcionamento do sistema, baseada em redes de filas estendidas (Soares, 1990). Tal modelagem tem o objetivo de relacionar os principais recursos de processamento e comunicação envolvidos sob

o gerenciamento de um sistema operacional e expor as dependências existentes entre eles.

O Capítulo 3 apresenta a arquitetura proposta, mostrando como os frameworks descritos em (Gomes, 1999) podem ser especializados ou mesmo estendidos para a definição de um modelo adaptável aos diversos mecanismos de provisão de QoS em sistemas operacionais. Acompanhando a descrição dos componentes de cada framework são mostrados exemplos de sua instanciação para um contexto específico.

Em seguida, o Capítulo 4 propõe um cenário real de uso da arquitetura, sobre o sistema operacional Linux. Além da instanciação dos frameworks definidos no Capítulo 3, são descritas as modificações e configurações necessárias sobre o sistema, para que sejam oferecidos alguns mecanismos para adaptabilidade e provisão de QoS.

Por fim, no Capítulo 5 são feitas as considerações finais sobre o trabalho, destacadas as contribuições da dissertação e relacionados os possíveis trabalhos futuros.