Douglas José Soares Rodrigues – ra011104

Luiz André Barroso, Jeffrey Dean, Urs Hölzle
Web Search for a Planet: The Google Cluster Architecture
pp. 119-129 – IEEE Micro 2003

Poucas aplicações demandam tanta computação por requisição quanto os buscadores de internet. A abordagem do Google para a solução deste problema é a utilização de um cluster de mais de 15 000 PCs, executando software tolerante a falhas. Os mais importantes fatores que influenciam esta abordagem são o uso eficiente de energia e a taxa preço/desempenho. Neste ambiente, o consumo de energia e a refrigeração são fatores muito significativos.

O objetivo da arquitetura é prover a confiabilidade da aplicação no software, viabilizando a utilização de PCs comuns com elevado poder computacional, a um baixo custo. A confiabilidade da aplicação se deve à distribuição do serviço entre várias máquinas, e a detecção automática de falhas. Os computadores do cluster estão distribuídos em vários sites, e um mecanismo de load-balancing é responsável pela distribuição das requisições.

A execução da consulta consiste em dois passos: primeiro, as palavras que compõe a busca são localizadas em índices invertidos, que mapeiam uma palavra em uma lista de documentos; depois as listas de documentos são unificadas para ordenar os resultados. Este processo é altamente paralelizável, a partir da divisão do índice em vários pedaços, cada um com um subconjunto de documentos escolhido aleatoriamente.

Os racks do Google são compostos por 40 processadores 80x86. Diferentes gerações de processadores compõem os racks, desde o Celeron de 533 MHz ao Pentium II de 1.4 GHz. Cada servidor contém um disco IDE de 80 Gb. Os servidores em um mesmo rack se comunicam por uma interface Ethernet de 100-Mbps, enquanto os racks são ligados a um switch gigabit. O critério de seleção dos processadores é o custo por consulta, na forma da soma dos gastos (incluindo a depreciação) e os custos operacionais dividida pelo desempenho.

A tabela abaixo mostra algumas estatísticas dos servidores de índices

Característica	***Valor***
Ciclos por instrução	1.1
Branch mispredict	5%
Misses de instruções no cache L1	0.4%
Misses de dados no cache L1	0.7%
Misses no cache L2	0.3%

A aplicação apresenta um CPI razoavelmente alto, considerando que o Pentium III é capaz de despachar três instruções por ciclo. A mesma carga executando num Pentium 4 apresenta aproximadamente o dobro do CPI, e a mesma taxa de "branch mispredict", mesmo podendo fazer o despacho de mais instruções paralelamente, e tendo uma lógica de predição de branches mais elaborada. De fato, não há muito paralelismo em nível de instruções a ser explorado neste tipo de aplicação.

É observado um bom desempenho para o cache de instruções, reflexo do código de loop pequeno das consultas. Os dados do índice não se beneficiam de localidade temporal, devido à imprevisibilidade das buscas, entretanto a localidade espacial é altamente explorada, onde o pre-fetching do hardware (ou linhas de cache maiores) apresenta um desempenho melhor.