Redes de Alto-Desempenho

Prof. Mario Dantas
mardantas@computer.org

Bibliografia

1 - *Tecnologias de Redes de Comunicação e Computadores,* ***Mario Dantas,*** Axcel Books, ISBN 85-7323-169-6

2- *High Performance Networks - Technology and Protocols* (Editor)**Ahmed N. Tantawy**, Kluwer, ISBN 0-7923-9371-6;

3 - *Sharing Bandwidth*, **Simon St. Laurent**, IDG Books, ISBN 0-7645-7009-9;

4 - *Internetworking with TCP/IP - Volume I Principles, Protocols, and Architecture,* **Douglas E. Comer**, Third Edition, Prentice Hall, 1995, ISBN 0-13-216987-8;

Bibliografia

4 - *TCP/IP and NFS - Internetworking in a Unix Environment,* ***Michael Santifaller****, Addison-Wesley, 1991, ISBN 0-201-54432-6;*

5 - Papers :
- *The Virtual Interface Architecture,* **Dave Dunning et al**, IEEE Micro, pp. 66-76, March-April 1998.
- *Myrinet: A Gigabit-per Second Local Area Network,* **N. Boden et al,** IEEE Micro, pp 29-36, Feb 1995

Conteúdo Programático

(I) Arquitetura de Redes e Internet :

- O modelo ISO/OSI e a arquitetura TCP/IP;
- Estratégia de compartilhamento de enlaces;
- Monitoração e melhor utilização da largura de banda.

Conteúdo Programático

(II) Tecnologias de Redes de GIGABIT

- Protocolos MAC para alta velocidade;
- Arquitetura de redes óticas;
- Fibre Channel;
- HIPPI;
- VIA;
- Myrinet.

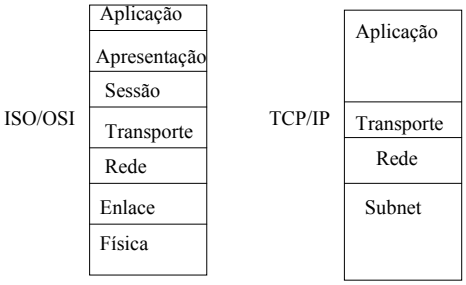
Conteúdo Programático

(III) Protocolos de Alto-Desempenho :

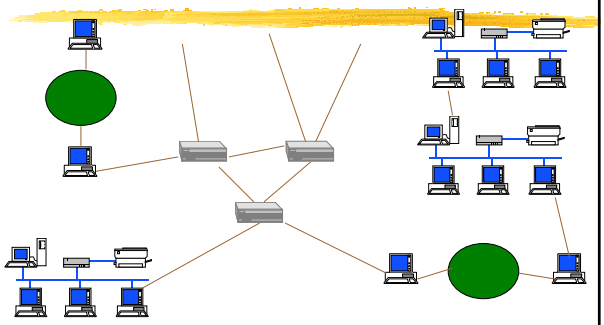
- Protocolos *lightweight* para redes de alta velocidade;
- Revisão de implementação de protocolos para alto desempenho.
- Resultados Experimentais

(I) Arquitetura de Redes e Internet

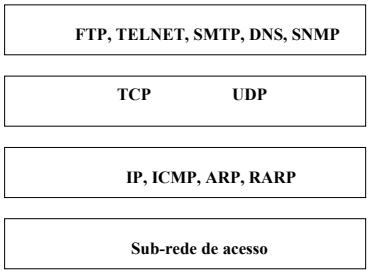
(1.1) - O Modelo ISO/OSI e a Arquitetura TCP/IP



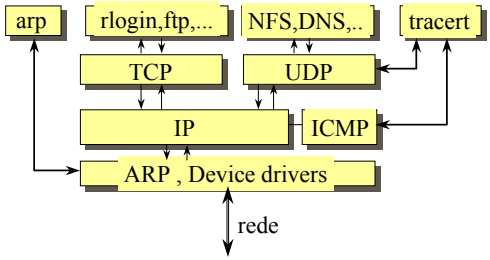
Internet



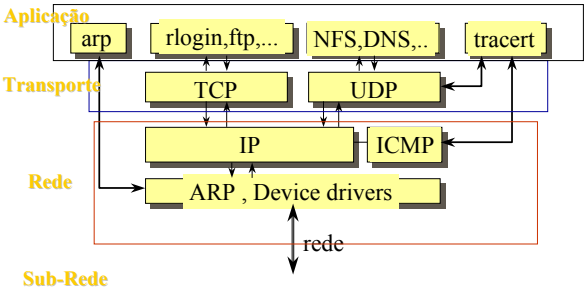
Arquitetura Internet



Família de Protocolos TCP/IP



Família de Protocolos TCP/IP

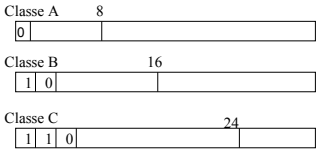


Sub-Rede de Acesso

- Ethernet, Token Ring, Token Bus ;
- FDDI, CDDI;
- X.25, Frame Relay;
- MTU (Maximum Transmission Unit).

IP (Internet Protocol)

- Endereçamento IP



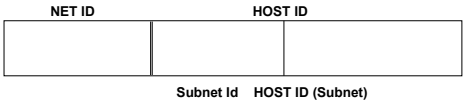
Endereçamento

- Classe A [0,126]
  - 0.1.0.0 (16.777.216 endereços de *hosts*)
- Classe B [128,191]
  - 164.41.14.0.0 (65.536 endereços de *hosts*)
- Classe C [192,223]
  - 196.25.15.0 (256 endereços de *hosts*)

Endereços Especiais

- Os campos Net Id e Host Id possuem significados diferentes quando possuem todos seus bits em zero (0) ou em um (1)
- Todos bits em um significa broadcast
  - Net Id: para todas as redes
  - Host Id: para todos os hosts dentro da rede
  - ex.: 192.31.235.255
- Todos bits em zero significa esta rede ou este host
  - ex.: 0.0.0.10
- LoopBack Address
  - 127.0.0.0

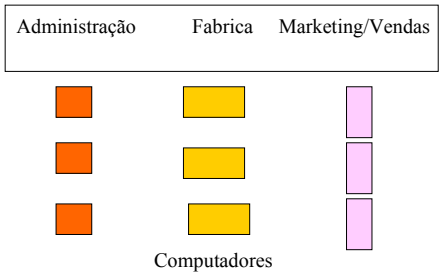
Sub-redes (Subnets)



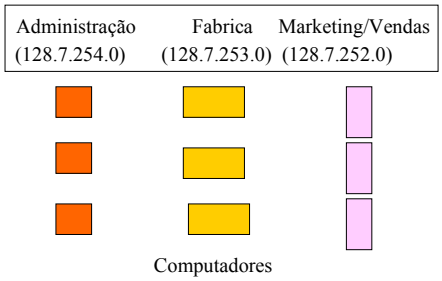
Máscara (Mask): usado para determinar o Net Id e o Host Id do endereço. Os bits em um (1) representam a parte do Net Id e Subnet Id, equanto que bits em zero (0) representam o Host Id (Subnet)

ex.: Classe B 143.54.0.0
Sub-rede A: 143.54.10.0
Sub-rede B: 143.54.20.0
Máscara: 255.255.255.0

Indústria X



Indústria X (128.7.0.0)



Indústria X
(128.7.0.0)

Administração (128.7.254.0)	Fabrica (128.7.253.0)	Marketing/Vendas (128.7.252.0)
128.7.254.12 Lúcia	128.7.253.4 Mario	128.7.252.55 John
128.7.254.56 João	128.7.253.1 Lula	128.7.252.65 Carol
128.7.254.64 Maria	128.7.253.2 Libanês	128.7.252.89 Catherin
	Computadores	

Classe de END.	Máscara Default (Binária)	Máscara Default (Decimal)
A	11111111.00000000.00000000.00000000	255.0.0.0
B	11111111.11111111.00000000.00000000	255.255.0.0
C	11111111.11111111.11111111.00000000	255.255.255.0

Como funcionam as máscaras da Subrede ?

As máscaras são calculadas através da operação do AND lógico sobre endereços envolvidos na rede que desejam se comunicar. Assim, observe o exemplo dos endereços.

Caso 1 - Considere o endereço da funcionária Lúcia, 128.7.254.12. Como a máscara para uma rede da classe B é 255.255.0.0, teríamos a seguinte operação :

```
11111111.11111111.00000000.00000000
10000000.00000111.11111110.00001100
-----------------------------------------------------
10000000.00000111.00000000.00000000   (128.7.0.0)
```

Como funcionam as máscaras da Subrede ?

Caso 2 - Considere o endereço da funcionário Mario, 128.7.253.4. Como a máscara para uma rede da classe B é 255.255.0.0, teríamos a seguinte operação :

```
11111111.11111111.00000000.00000000
10000000.00000111.11111101.00000100
-----------------------------------------------------
10000000.00000111.00000000.00000000   (128.7.0.0)
```

Como funcionam as máscaras da Subrede ?

Caso 3 - Considere o endereço da funcionária Catherin, 128.7.252.89. Como a máscara para uma rede da classe B é 255.255.0.0, teríamos a seguinte operação :

```
11111111.11111111.00000000.00000000
10000000.00000111.11111100.01011001
-----------------------------------------------------
10000000.00000111.00000000.00000000   (128.7.0.0)
```

Como funcionam as máscaras da Subrede ?

Como resolver este problema ?

Como funcionam as máscaras da Subrede ?

Caso 1 - Considere o endereço da funcionária Lúcia, 128.7.254.12.
Com a máscara correta da rede 255.255.255.0, teríamos
a seguinte operação :

11111111.11111111.11111111.00000000
10000000.00000111.11111110.00001100

10000000.00000111.00000000.00000000 (128.7.254.0)

Como funcionam as máscaras da Subrede ?

Caso 2 - Considere o endereço da funcionário Mario, 128.7.253.4.
Com a máscara correta da rede 255.255.255.0, teríamos
a seguinte operação :

11111111.11111111.11111111.00000000
10000000.00000111.11111101.00000100

10000000.00000111.00000000.00000000 (128.7.253.0)

Como funcionam as máscaras da Subrede ?

Caso 3 - Considere o endereço da funcionária Catherin, 128.7.252.89.
Com a máscara correta da rede 255.255.255.0, teríamos
a seguinte operação :

11111111.11111111.11111111.00000000
10000000.00000111.11111100.01011001

10000000.00000111.00000000.00000000 (128.7.252.0)

Exemplo Sub-rede

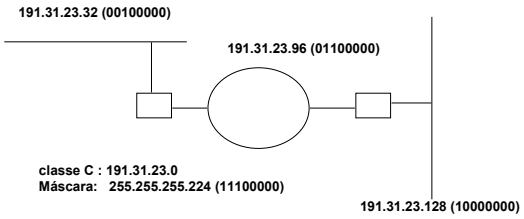
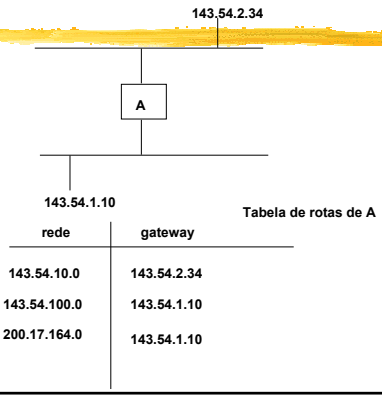


Tabela de Rotas



IP

Vers	HLEN	Service Type	Total Legth	
Identification			Flags	Offset
TTL	Protocol		CheckSum	
Source IP				
Destination IP				
Options				PAD
DADOS				

Campos IP

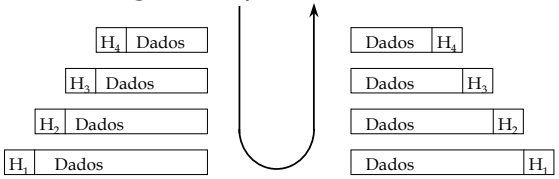
- **Vers**: versão do IP utilizada. Versão autal é a 4
- **Hlen:** tamnha do cabeçalho do datagrama
- **Service Type**: especifica qual a forma de se lidar com o datagrama. Possui 8 bits que indicam os seguinte requisitos:
  - Precedência
  - Mínimo de atraso na transmissão
  - Alto Throughput
  - Alta confiabilidade
- **Total Len**: tamanho total do datagrama

Encapsulamento do Datagrama

- Os datagramas podem ser fragmentados devido ao MTU da sub-rede
- Tamanho máximo de 65531 octetos
- Os campos Identification, Flags e Fragment Offset são usados na fragmentação

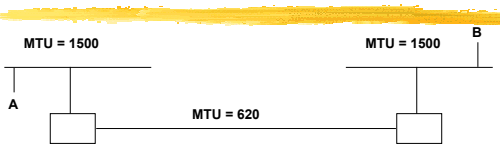
Encapsulamento do Datagrama

- Quebra em pacotes
- Tráfego de sequência de bits



H_x - Cabeçalho (Header) - Controle
Dados - Não tratado pelo nível x

Encapsulamento do Datagrama



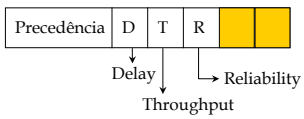
Datagrama de 1400 octetos : A -> B

	Identification	Flags	Offs
Frag1:	xxxx	010	0
Frag2:	xxxx	010	600
Frag3:	xxxx	001	1200

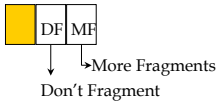
- Fragmentos só são remontados no host destino

Datagrama IP

Tipo de Serviço



Flags



Campos IP

- TTL (Time To Live): número máximo de gateways que um datagrama pode passar. Cada gateway ao repassar um datagrama decrementa de um este valor, caso resulte em zero, o datagrama é descartado
- Protocol: tipo do protocolo encapsulado no datagrama
- CheckSum: garante a integridade dos dados
- Source IP: endereço da máquina emissora do datagrama
- Destination IP: endereço da máquina destino do datagrama

Campo de Opções

- Registro de rota
- Especificação de rota
- Tempo de Processamento de cada gateway
- ...
  - Opção COPY/Fragmentação

Processamento no Roteador

- Se o roteador nao tem memória suficiente, o datagrama é descartado
- Verificação do Checksum, versão, tamanhos
  - O Checksum é recalculado, se for diferente do datagrama, este é descartado
- Decremento do TTL
  - se zero, o datagrama é descartado

Processamento no Roteador

- Pode -se considerar o campo Service Type
- Se for necessário e permitido, o datagrama pode ser fragmentado. Cria-se um cabeçalho para cada fragmento, copiando as opções, aplicando o novo TTL e o novo Checksum
- Tratamento do campo opção
- Repasse para a sub-rede destino

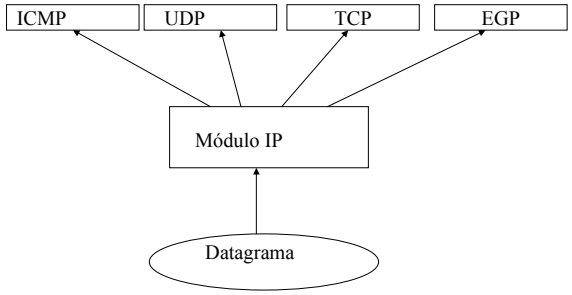
Processsamento no Host Destino

- Verificação do Checksum, versão, tamanhos
  - O Checksum é recalculado, se for diferente do datagrama, este é descartado
- Se o datagrama é fragmentado, é disparado um temporizador que evitará o espera indefinida dos outros fragmentos do datagrama original
- Entrega do campo de dados do datagrama para o processo indicado no campo Protocol

Recursos Críticos para o Desempenho IP

- Largura de banda disponível
- Memória disponível para buffers
- Processamento da CPU

Demultiplexação na camada de rede



Protocolo IP - Multiplexação

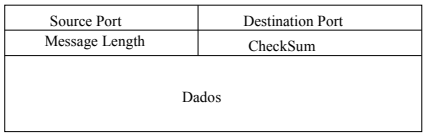
- Convenção para auto identificação dos datagramas

Protocol	Serviço
0	IP - pseudo
1	ICMP
6	TCP
17	UDP

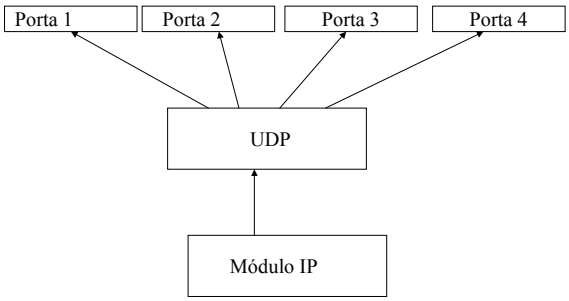
UDP (User Datagram Protocol)

- Protocolo de transporte não orientado à conexão
- Não implementa nenhum mecanismo de recuperação de erros
- São identificados os processos origem e destino através do conceito de porta
- O campo de CheckSum é opcional

UDP



Demultiplexação na camada de Transporte



TCP (Transmission Control Protocol)

- Protocolo de transporte orientado à conexão
- Implementa mecanismos de recuperação de erros
- Usa o conceito de porta
- Protocolo orientado a stream

TCP (Transmission Control Protocol)

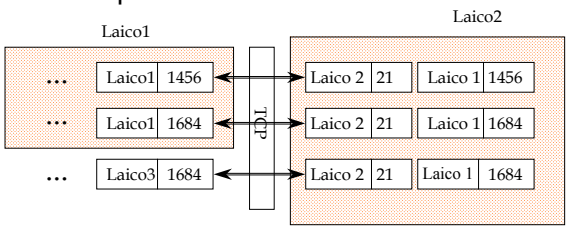
- Usado para aplicações cliente-servidor, serviços críticos,...
- Faz a multiplexação de mensagens para as aplicações
- Conexão (IP,port) <--> (IP,port)
  Permite multiplas sessões do mesmo serviço

Janela Deslizante

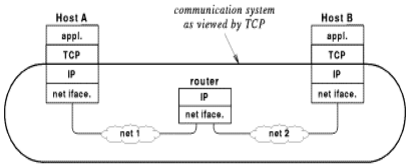
- Cada octeto é numerado
- O tamanho da janela determina o número de octetos que podem ser transmitidos sem reconhecimento
- Através do mecanismo de PIGGYBACK pode-se reconhecer um bloco de octetos via um segmento de dados

TCP - Exemplo

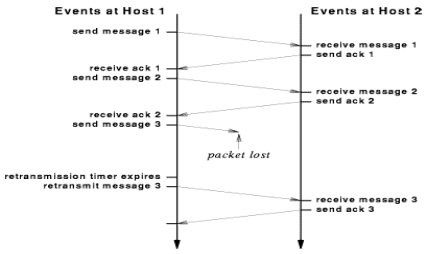
- FTP - port 21



TCP - Visão de comunicação

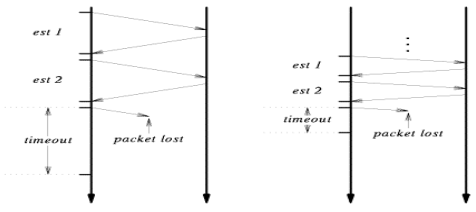


TCP - Confiabilidade

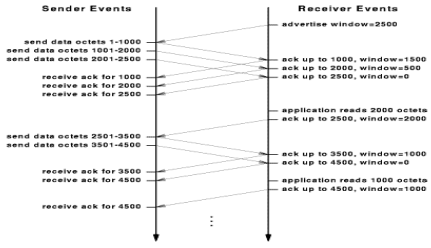


TCP - timeout adaptativo

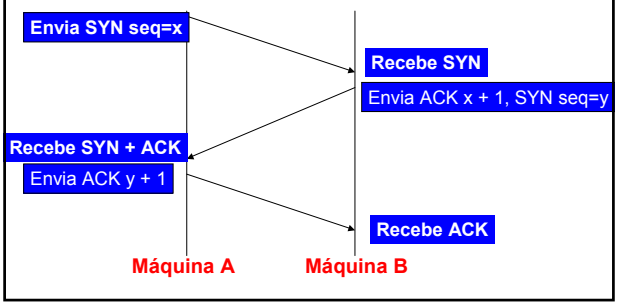
- Verifica o tempo de resposta e ajusta o timeout



TCP - Janela deslizante

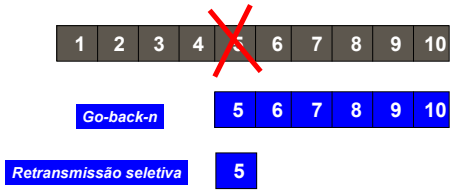


TCP - Estabelecimento de Conexão



TCP - Reconhecimento e retransmissão

Go-back-n x Retransmissão seletiva



Diferenças entre IPv4 e IPv6

Introdução

O protocolo IPv6 mantêm muitas das funções do IPv4, e que foram responsáveis pelo sucesso da Internet. Alguns exemplos destas características são :

- entrega de datagrama não confiável;
- permite que o remetente escolha o tamanho do datagrama;
- requer que o número de *hops* seja estabelecido no remetente.

- Apesar da similaridade ser verificada em vários pontos nos protocolos IPv4 e IPv6, o novo protocolo adota mudanças significativas.

Exemplo de modificações consideradas no IPv6 são o tamanho de endereçamento e algumas facilidades adicionais (como uma maior flexibilidade para o uso da QoS).

As modificações que foram implementadas no IPv6 podem ser agrupadas em cinco grandes grupos :

- Maior endereço : é uma das características mais marcante da nova versão do IP. O IPv6 quadruplicou o tamanho do IPv4 de 32 bits para 128 bits (*esta expansão foi projetada para atender um futuro ainda não imaginado)* ;

- Formato de cabeçalho flexível : o formato do datagrama IP não é compatível com o antigo datagrama IPv4. A abordagem foi implementar um formato com uma série de cabeçalhos adicionais. Em outras palavras, diferente do IPv4 onde é usado um tamanho fixo para o formato onde os campos têm tamanhos fixos;

- Opções melhoradas : as opções disponíveis no IPv6 são mais poderosas quando comparadas com o IPv4;

• Suporte para reserva de recursos : existe um mecanismo que permite a alocação prévia de recursos da rede. Desta forma, aplicações como vídeo em tempo real que necessitam de um garantia de largura de banda e baixo retardo podem ser atendidas;

• Previsão para uma extensão do protocolo : esta característica é tida por muitos como a maior melhoria. Em outras palavras, existe uma flexibilidade de expansão do protocolo para novas realidades. De forma oposta ao IPv4, onde uma existe uma especificação fechada e *completa* do protocolo.

ENDEREÇAMENTO IPv6

O endereço do protocolo IPv6 é composto de 16 octetos. A idéia do tamanho do endereço pode ser mensurada pela afirmação :

Cada pessoa no globo poderá ter suficiente endereço para ter a sua própria Internet, tão grande quanto a Internet atual.

Um endereçamento de 16 octetos representam 2 elevado a 128 valores válidos de endereços. Em outras palavras, o tamanho de endereços é da ordem de 3.4 x (10 elevado a 38).

Se os endereços fosse alocados a uma taxa de um milhão de endereços a cada micro-segundo, teríamos que ter vinte anos para que todos os endereços fossem alocados.

Notação do Endereço IPv6

Devido a dificuldade dos humanos de trabalhar com endereços binários, e grandes, o grupo responsável pelo endereçamento do IPv6 imaginou uma nova notação, ***colon hex***.

O que vem a ser isto ?

Colon Hex

é uma representação caracterizada por valores hexadecimais separados por dois pontos. Assim teríamos para o exemplo genérico a seguir:

Decimal

255.255.10.150.128.17.0.0.255.255.255.255.100.140.230.104

Colon Hex

FFFF:A96:8011:0:FFFF:FFFF:648C:E668

Colon Hex

A notação tem claras vantagens, por solicitar menos dígitos e menos separadores. Outras duas vantagens desta abordagem são seguintes técnicas:

• *Compressão de zeros* - em exemplo pode ser o endereço

FF05:0:0:0:0:0:0:B3

Este endereço pode ser representado como :

FF05::B3

Esta técnica somente pode ser usada uma vez num endereço.

Colon Hex

• A abordagem de *colon hex* permite que adotemos a sintaxe de sufixo decimal com ponto. O objetivo é a manutenção da compatibilidade de transição entre o IPv4 e o IPv6.

Exemplo :

0:0:0:0:0:0:0:128.10.2.1

Como seria o uso da técnica de compressão de zeros para este endereço ?

Como seria o uso da técnica de compressão de zeros para este endereço ?

Resposta

:: 128.10.2.1

Tipos de Endereços Básicos do IPv6

Os endereços de destino num datagrama no IPv6 tem as seguintes três categorias:

• *Unicast* - o endereço especifica um *host simples (computador ou roteador)* que o datagrama deverá ser enviado pelo menor caminho ;

• *Cluster* - o destino é um conjunto de computadores que compartilham um único prefixo de endereço (todos conectados a uma mesma rede física). O datagrama deverá ser encaminhado para o grupo e entregue a um membro do grupo (o mais perto possível).

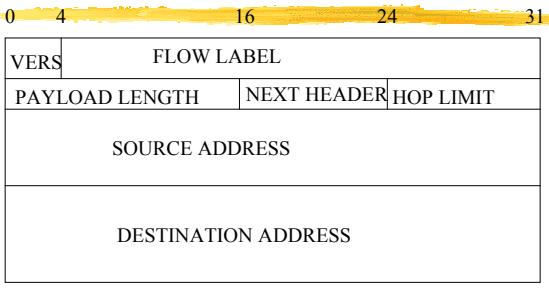
• *Multicast* - o destino é um conjunto de computadores, possivelmente em locais diferentes. Assim o datagrama deverá ser entregue para cada membro do grupo multicast usando facilidade de multicast de hardware, ou broadcast se possível.

Forma Geral do Datagrama IPv6

Base Header	Extension Header 1	...	Extension Header n	Data

A figura acima ilustra uma forma geral um datagrama IPv6 genérico. Somente o campo do cabeçalho básico é necessário, os demais são opcionais.

O formato do IPv6 base header é ilustrado abaixo :



FLOW LABEL

Este campo permite que os pacotes que tenham que ter um tratamento diferenciado sejam assim tratados.

O campo tem tamanho de 20 bits, composto do endereço de origem e IP destino, permitindo que os roteadores mantenham o estado durante o fluxo ao invés de estimar a cada novo pacote.

As aplicações são obrigadas a gerar um *flow label* a cada nova requisição. O reuso do *flow label* é permitido quando um fluxo já está terminado ou foi fechado.

FLOW LABEL

A utilização de campo *flow label* prove aos roteadores uma maneira fácil de manter as conexões e manter o fluxo de tráfego numa mesma taxa.

PRIORITY

A utilização do campo *priority* prove aos programas a facilidade de identificar a necessidade de tráfego que os estes necessitam. O uso efetivo, ou normalização, de como este campo junto com o *flow label* devem plenamente operar ainda estão em discussão.

O campo de 8-bits destinado a *Classe* está no momento a nível de desenvolvimento. Todavia, os 4 bits de prioridade podem nos ajuda a entender o que poderemos ter pela frente.

IPv6 Priority Values

Valor	Descrição
0	Tráfego sem características
1	Tráfego Filler (fluxo contínuo de informação onde o tempo particularmente não interessa)
2	Transferência de dados sem supervisão (ex. e-mail)
3	Reservado
4	Transferência grande de dados com supervisão (ex. HTTP, FTP e tráfego NFS).

IPv6 Priority Values

Valor	Descrição
5	Reservado
6	Tráfego Interativo (ex. telnet)
7	Tráfego de Controle da Internet (informação usada por dispositivos que fazem parte da Internet, como roteadores, switches e dispositivos que empregam o SNMP para reportar estados).
8-15	Pacotes em processos que não podem controlar congestionamentos. Pacotes com valor 8 serão descartados antes do de valor 15.

IPv6 fragment extension header é ilustrado

0	8	16	29	31
N. Header	Reserved	Fragment Offset		MF
Identificação do Datagrama				

Distribuição de Endereços IPv6

A forma de distribuição dos endereços têm gerado muitas discussões. As discussões ficam baseadas em dois pontos principais :

- Como fazer a gerência de distribuição dos endereços ?
- Como mapear um endereço para um destino ?

Como fazer a gerência de distribuição dos endereços ?

Esta discussão é baseada em qual autoridade deve ser criada para gerenciar a distribuição de endereços.

Na Internet atual temos dois níveis de hierarquia. Em outras palavras, temos um primeiro nível que é responsabilidade da autoridade da Internet. No segundo nível é responsabilidade da organização.

O IPv6 permite múltiplos níveis. Existe uma proposta em tipos de níveis do IPv6 semelhante ao IPv4.

Como fazer a gerência de distribuição dos endereços ?

010	Provider ID	Subscriber ID	Subnet ID	Node ID

010 - tipo de endereço, no caso 010 é um endereço que diz o tipo de provedor auferido;
Provider ID - identificação do provedor
Subscriber ID - identificador do assinante
Subnet ID - informação da rede do assinante
Node ID - informação sobre um nó do assinante.

Como mapear um endereço para um destino ?

Esta pergunta deve ser respondida com o desempenho com meta. De um outra forma, a eficiência computacional deverá ser levada em conta. Independente de autoridades na rede, um datagrama deverá ser analisado e os melhores caminhos deverão ser escolhidos.

TCP - Parâmetros

- MSS - Maximum segment Size
- Padronização de ports
- Controle de Congestionamento

Port	Serviço
15	netstat
21	ftp
23	telnet
25	smtp (mail)

TCP

Source Port			Destination Port
Sequence Number			
Ack Number			
Hlen	Reserved	Cod Bits	Window
CheckSum			Urgent Pointer
Opções			
Dados			

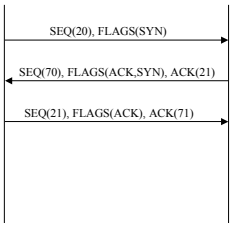
Campos do TCP

- **Portas**: identificam os processos origem e destino
- **Número de sequência**: número do primeiro octeto do campo de dados
- **Número do ACK**: número do octeto que é esperado pelo destino, sendo todos os octetos de número inferior reconhecidos
- **HLEN**: Tamanho em bytes do cabeçalho TCP
- Bytes de código: URG, ACK, PSH, RST, SYN, FIN
- **Window**: Tamanho em octetos da janela que é aceito pelo emissor

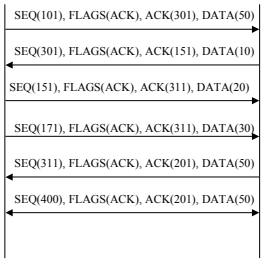
Campos do TCP

- **Checksum**: verificador de erros de transmissão
- **Urgent pointer**: fornece a posição dos dados urgentes dentro do campo de dados
- **Opção**: pode conter negociação de opções tal como o MSS (Maximum Segment Size)

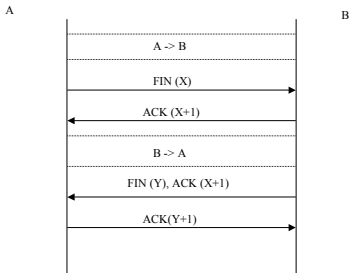
Estabelecimento de Conexão



Troca de Dados



Liberação de Conexão



Porta Padrão

- FTP (21)
- FTP-DATA (20)
- TELNET (23)
- SMTP (25)
- NNTP (119)

ICMP (Internet Control Message Protocol)

- Oferece funções de gerência
- Encapsulado no datagrama IP
- Emitido por um gateway intermediário oupelo host destino

ICMP

Type	Code ->
Code	CheckSum ->
CheckSum	Information

Mensagens ICMP

- Echo Request / Echo Reply
- Destination Unreachable
- Source Quench
- Redirect
- Time Exceeded for a Datagram
- Parameter Problem on a Datagram
- Timestamp Request / Timestamp Reply
- Address Mask Request / Address Mask Reply

Destination Unreachable

- Mensagem aplicada nos diversos casos em que o datagrama não pode ser entregue ao destinatário especificado
- Causas (Código): sub-rede inacessível, estação inacessível, protocolo inacessível, fragmentação necessária e campo flag não configurado, falha na rota especificada, repasse proibido por um filtro no gateway intermediário

Time Exceeded

- Retornado por um gateway quando o TTL de um datagrama expira.
- Retornado por um host quando o tempo para remontagem de um datagrama fragmentado expira

Parameter Problem

- Relata a ocorrência de erros de sintaxe ou semântica no cabeçalho de datagrama
- No campo código tem-se o ponteiro para o campo do cabeçalho que gerou o erro contido no campo de dados da mensagem ICMP

Source Quench

- Serve como regulador de fluxo de recepção. Essa mensagem é gerada por gateways ou hosts quando eles precisam reduzir a taxa de envio de datagramas do emissor. Ao recebê-la, o emissor deve reduzir a taxa de emissão.

Redirect

- Usada pelo gateway para notificar um host sobre uma rota mais adequada ao destinatário do datagrama por ele enviado. É usado pelos hosts para atualizar as suas tabelas de roteamento
- Neste tipo de mensagem não há descarte de datagramas

Echo Request e Echo Response

- São mensagens usadas para testar se a comunicação entre duas entidades é possível. O destinatário é obrigado a responder a mensagem de Eco com a mensagem de Resposta de Eco
- Usado para estimar o throughput e o round trip time
- ping -s laicox.cic.unb.br

Timestamp Request e Timestamp Response

- São usadas para medir as características de atraso no transporte de datagramas em uma rede Internet. O emissor registra o momento de transmissão na mensagem. O destinatário, ao receber a mensagem, faz o mesmo.
- No campo de dados é enviado um identificador, número de seqüência, tempo de envio do request, tempo de recebimento do request e tempo envio do response
- Pode-se calcular o tempo de processamento do datagrama no host destino

Address Mask Request e Address Mask Response

- Utilizadas por uma estação na recuperação da máscara de endereços quando é aplicado o sub-endereçamento ao endereço IP. O host emite a mensagem e o gateway responsável responde com a descrição da máscara

ARP (Address Resolution Protocol)

- O nível IP utiliza para o transporte dos datagramas os gateways interligando as sub-redes. Para tal é necessário o conhecimento do endereço físico dos gateways.
- O ARP realiza o mapeamento do endereço lógico IP para o endereço físico da sub-rede
- O ARP é encapsulado direto no protocolo da sub-rede
- O host que pretende mapear um endereço IP para o físico deve enviar um pedido ARP no modo broadcast na rede. O host que receber e verificar que o endereço é o seu, responde com o seu endereço físico

ARP (Address Resolution Protocol)

- Cada host possui um cache dos mapeamentos realizados de forma a utilizar a busca do endereço físico. Contudo tais entradas armazenadas possuem um tempo de vida limitado, permitindo alterações nos endereços.
- Um host ao receber um pedido ARP pode atualizar o sua cache mesmo que o endereço procurado não seja seu

ARP (Address Resolution Protocol)

<table>
<tr><td colspan="2">Hard Type</td><td>Proto Type</td></tr>
<tr><td>Hlen</td><td>Plen</td><td>Operation</td></tr>
<tr><td colspan="3">Sender HA</td></tr>
<tr><td colspan="2">Sender HA</td><td>Sender IP</td></tr>
<tr><td colspan="2">Sender IP</td><td>Target HA</td></tr>
<tr><td colspan="3">Target HA</td></tr>
<tr><td colspan="3">Target IP</td></tr>
</table>

RARP (Reverse Address Resolution Protocol)

- Usado por um host para descobrir o seu endereço lógico IP a a partir do seu endereço físico (Diskless)
- O RARP é encapsulado diretamente no protocolo da sub-rede
- Quando o host necessita do seu endereço IP envia um pedido no modo broadcast. O servidor RARP que mantém uma tabela de mapeamento responde.
- Utiliza o mesmo formato de protocolo do ARP
- Pode haver mais de um servidor RARP

Aplicação Internet

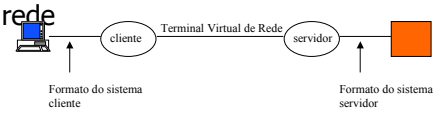
- Modelo Cliente-Servidor
  - Servidor é um processo que espera pedidos de conexão (se TCP), aceita requisições de serviço e retorna uma resposta
  - Cliente é um processo que inicia uma conexão (se TCP), envia requisições e espera resposta. Geralmente possui uma interface com o usuário
  - A comunicação entre o cliente e o servidor acontece através de um protocolo de aplicação
  - Contra-exemplo: emulador de terminal

Aplicações Internet

- Telnet
- Ftp (File Transfer Protocol)
- SMTP (Simple Mail Transfer Protocol)
- DNS (Domain Name System)
- Gopher
- Http (Hypertext Transfer Protocol)

Telnet

- Implementa o serviço de terminal remoto
- Utiliza o conceito de terminal virtual de rede



Telnet

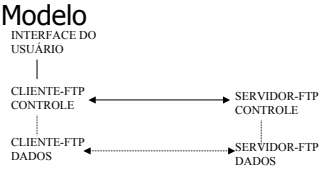
- Utiliza TCP/IP
- Terminal Virtual de Rede
  - Usa ASCII com 7 bits para representar dados e 8 bits para representar comandos
  - Interpreta caracteres de controle ASCII, tais como CR e LF
  - Usa o caractere IAC (Interpreter for command - 255) para referenciar um comando, tal como 255 244 para interromper um programa na máquina remota

Telnet

- Possibilita a negociação simétrica de opções com os comandos WILL, DO, DON'T, WON'T
  - Opções disponíveis:
    - Eco
    - Tipo de terminal
    - Uso de EOR (End of Register)
    - Transmissão de dados com 8 bits

FTP

- Protocolo que permite a transferência de arquivos entre computadores na Internet
- Usa TCP/IP
- Modelo



FTP

- Tipos de dado: ASCII, EBCDIC, Imagem, Local
- Estrutura do arquivo: Não-estruturado, Orientado a registro, Paginado
- Modos de Transferência: Fluxo Contínuo, Blocado, Comprimido
- Re-início de Transferência
- Comandos: controle de acesso, manipulação de diretórios, parâmetros de transferência e de serviços

Controle de Acesso

- user
- pass
- acct
- smnt
- rein
- quit

Manipulação de Diretório

- cwd (Change Working Group)
- cdup (Change to Parent Directory)
- mkd (Make Directory)
- rmd (Remove Directory)
- pwd (Print Working Directory)
- list
- nlst (Name List)

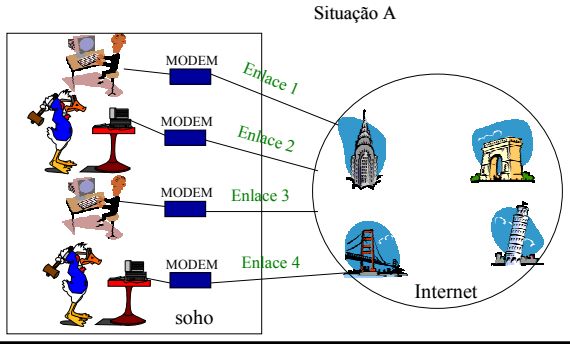
(I) Arquitetura de Redes e Internet

(1.2) - Estratégia de compartilhamento de enlaces

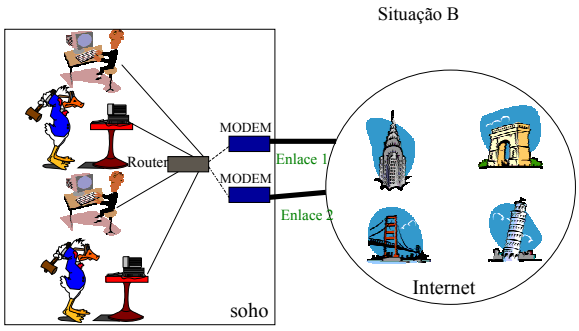
• Acesso via 

• Qual seria então a  para melhorar o compartilhamento ?

Estratégia de compartilhamento de enlaces



Estratégia de compartilhamento de enlaces



Estratégia de compartilhamento de enlaces

Definições :

- POTS (Plain Old Telephone Service) - linha de telefone convencional;
- ISDN (Integrated Services Digital Network)

Comentários :

1 - A comunicação nos *POTS* é analógica, a velocidade real máxima é de 53 Kbps (embora tecnicamente pode-se atingir 56Kbps);
2 - A comunicação nos *ISDN* é digital, onde a *BRI (Basic Rate Interface)* prove dois canais de dados do tipo B (de 56 Kbps ou 64Kbps) e um canal D (16 Kbps) de controle e discagem.

Estratégia de compartilhamento de enlaces

Comentários :

3 - O ISDN algumas vezes é conhecido como *It Still Does Nothing*. A razões são a falta de padronização de equipamentos, as instalações são caras e com (ainda) altas taxas de erro, não existe uma cobertura grande de regiões (mesmo na Califórnia).

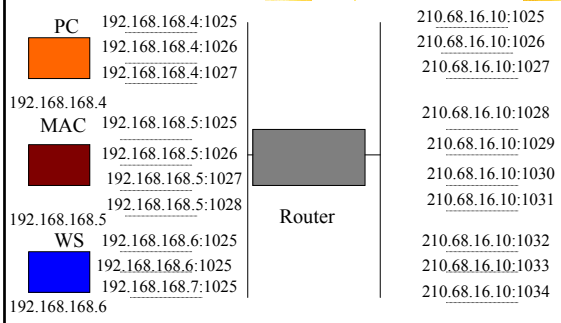
4 - Os serviços POTS são relativamente baratos, fazendo com que muitas soluções de conexões sejam baseadas nesta tecnologia.

Roteamento e Tradução de Endereços

Roteamento - permite que um dispositivo, ou processo, enviar informação entre *hosts* numa rede.

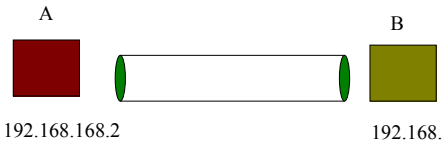
Tradução de Endereços - permite que uma rede *pequena* seja vista pela *Internet* como um simples nó. Desta forma, um único endereço IP (ou um conjunto pequeno de endereços) poderá ser compatilhado pela rede. Este aspecto *pode* prover dentre outras características uma proteção contra intrusos na rede.

Roteamento e Tradução de Endereços



Técnicas de Compartilhamento de Links

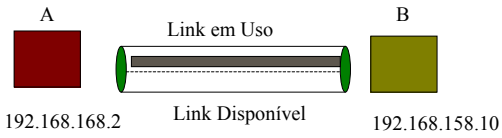
PPP (Point-to-Point Protocol) - este protocolo provê um mecanismo padrão para que dispositivos numa rede TCP/IP possam se comunicar através de uma linha de comunicação simples (ou única).



Técnicas de Compartilhamento de Links

Multilink PPP - A RFC 1717 define o *multilink* para o protocolo PPP visando que o mesmo possa permitir vários *links* simultâneos num único canal de comunicação. As ligações podem ser adicionadas, ou cortadas, visando a utilização com eficiência do canal.

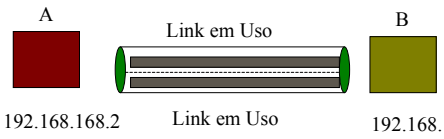
Exemplo - Tráfego Leve



Técnicas de Compartilhamento de Links

Multilink PPP - A RFC 1717 define o *multilink* para o protocolo PPP visando que o mesmo possa permitir vários *links* simultâneos num único canal de comunicação. As ligações podem ser adicionadas, ou cortadas, visando a utilização com eficiência do canal.

Exemplo - Tráfego Pesado



Técnicas de Compartilhamento de Links

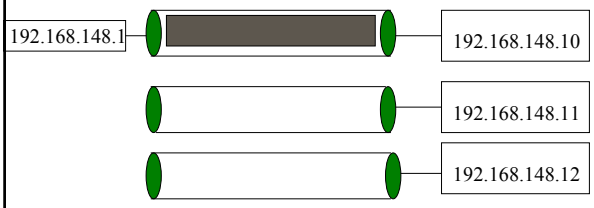
COLT (Connection Optimized Link Technologiy) - esta técnica foi criada pela empresa RAMP Networks para seus roteadores *(routers analógicos)*.

Embora RAMP empregue o PPP para seus protocolos seriais, eles não se utilizam do *Multilink PPP*. Ao contrário da criação de um *grande canal*, o COLT distribui o tráfego por multiplos diferentes canais. Como exemplo empregando diferentes canais dos modems:

- temos no primeiro o *download* de *mail ;*
- no segundo existe uma conexão HTTP;
- no terceiro uma conexão FTP;

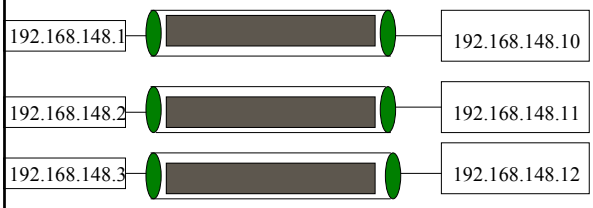
Técnicas de Compartilhamento de Links

COLT - Exemplo de Tráfego Leve



Técnicas de Compartilhamento de Links

COLT - Exemplo de Tráfego Pesado



Técnicas de Compartilhamento de Links

Software *vs* Appliances

Software - roteamento e tradução de endereços podem executar com muita facilidade num PC, ou Workstation. Com auxílio de pequenos pacotes de software comprados (exemplo - *iNet for Windows 95*, da Artisoft), ou freeware/shareware (exemplo - *IPMasq* para Unix).

Até mesmo o Multilink PPP pode já estar implemento no Sistema Operacional.

Técnicas de Compartilhamento de Links

Software *vs* Appliances

Appliance (dispositivo de uso específico) - a utilização destes dispositivos têm provado que :

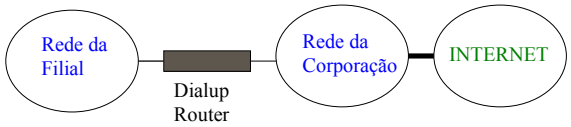
Vantagens:
- *mais baratos;*
- *mais fácil de gerenciar;*
- *menos suceptível a erros;*

Desvantagem:
- *falta de flexibilidade.*

Técnicas de Compartilhamento de Links

Roteadores Pequenos em Corporações

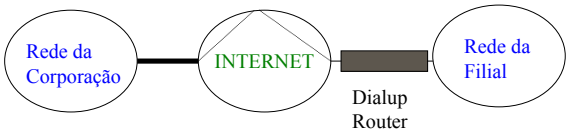
Situação A



Técnicas de Compartilhamento de Links

Roteadores Pequenos em Corporações

Situação B



Técnicas de Compartilhamento de Links

Soluções de Mercado

- *RAMP Networks - WebRamp 300e*
- *3COM - OfficeConnect Remote Dual Analog*
- *ISDN*
- *IP Masquerading/Linux*
- *GNAT Box*
- *IPNetRouter for Macintosh*

Técnicas de Compartilhamento de Links

RAMP Networks - WebRamp 300e

Este dispositivo é uma pequena *caixa* que inclui :

- 4 ports 10BaseT Ethernet Hub;
- 3 portas seriais para modem;
- 1 porta serial para console;
- luzes indicativas de *status* das portas;
- gerência através de um aplicativo WEB ou conexão telnet;
- prove o rotamento de IP e IPX;
- utiliza o protocolo COLT para três modem simultaneamente;
- pode usar o Multilink.

Técnicas de Compartilhamento de Links

OfficeConnect Remote Dual Analog

Este dispositivo cujo objetivo é a ligação de de pequenos escritórios com grande corporações tem as seguintes *features:*

- 4 portas Ethernet Hub;
- 2 portas de modem 56 Kbps;
- pacote de bridge e router para IP e IPX;
- Multilink PPP (*para criação de conexões de tamanho compatível com uma ligação ISDN*);
- CLI (Comand Line Interface) e Web-based adminstrative Interfaces;

Técnicas de Compartilhamento de Links

ISDN

Um grande quantidade de soluções utilizando a largura de banda ISDN existem. Muitas incorporam serviços de roteadores pequenos e tradução de endereços.

Uma solução ISDN, geralmente, é projetada mais do que uma simples substituição de linhas de modems. Muitas vezes a substituição é uma política da organização.

RAMP Networks, 3Com e Intel, dentre outras, comercializam pequenos roteadores ISDN.

Técnicas de Compartilhamento de Links

IP Masquerading/Linux

No mundo Unix, um exemplo é o Linux, estes sistemas operacionais proveêm a facilidade de tradução de endereços para redes externas. Um exemplo é a Internet.

Um servidor Linux, empregando o *IP Masquerading*, pode atuar como um roteador. Desta forma, o roteamento de mensagens e páginas WEB podem ser tratadas como se tivessemos um dispositivo dedicado.

Técnicas de Compartilhamento de Links

GNAT BOX

O *GNAT Box* é uma solução conhecida como *single-disk routing solution*. Através do suporte de uma rede interna, de uma rede externa eum serviço privado para Web Servers. Prove tradução de endereços (NAT - Network Address Translation).

É uma ferramenta pequena (*gnat*) que foi desenvolvida para ser instalada num PC com várias portas de rede, ou modem. O uso do PC exclusivo é um requerimento do aplicativo.

A arquitetura do máquina recomendada é : 16MB memória, processador 386 à Pentium xx, floppy disk e duas interfaces de rede. Não precisa de hard drive e caso exista este será ingnorado.

Técnicas de Compartilhamento de Links

GNAT BOX

O *GNAT* suporta até 16.384 conexões, as portas da rede externa aceitam até 100 endereços IP. O software pode empregar um DHCP próprio para as ligações internas/externas.

O GNAT pode ser avaliado antes de sua compra, todavia a cópia de avaliação só aceita 100 conexões concorrentes (www.gnatbox.com).

Técnicas de Compartilhamento de Links

IPNetRouter

O IPNetRouter é um protocolo de roteamento IP baseado no OpenTransport. O Protocolo possui a caracteristica de tradução de endereço (NAT) entre múltiplas interfaces de rede.

O OpenTransport controla apenas uma interface, enquanto as demais são gerenciadas pelo IPNetRouter. Através deste segundo pacote, o Machintosh pode ser considerando num ambiente de DNS e DHCP de outros fabricantes.

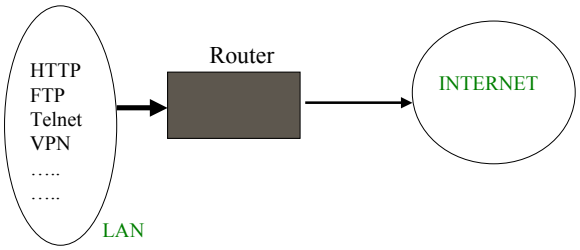
O IPNetRouter utiliza o PPP. Detalhes adicionais do pacote podem ser encontrados no endereço www.sustworks.com .

(I) Arquitetura de Redes e Internet

(1.3) Monitoração e melhor utilização da largura de banda.

PROBLEMA - A maioria das redes, a *Internet* é um bom exemplo, são caracterizadas pelo uso em larga escala (e de proposito geral) de transferência de diferentes tipos de informação (ex. conjunto de dados, e-mail, arquivos, aplicações multimídia, etc…)

Monitoração e melhor utilização da largura de banda



Monitoração e melhor utilização da largura de banda

PROBLEMA - O modelo TCP/IP não prove, originalmente, a priorização de tráfego e não garante que a entrega dos datagramas serão efetuadas num determinado tempo.

SOLUÇÃO (a) - Uma das possíveis soluções para o engarrafamento da informação que saem de um LAN é a adoção de uma política de *filtragem* do *tipo* e *volume* de tráfego que sai da sua LAN.

Monitoração e melhor utilização da largura de banda

PROBLEMA - O modelo TCP/IP não prove, originalmente, a priorização de tráfego e não garante que a entrega dos datagramas serão efetuadas num determinado tempo.

SOLUÇÃO (b) - Embora o IPv4 não tenha um conjunto de ferramentas adequadas para a gerência da rede, a existência de ferramentas IPv4 Network Management auxiliam na monitoração e melhor utilização da largura de banda.

Monitoração e melhor utilização da largura de banda

Quality of Service (QoS) - para que possamos ter os requisitos de QoS atendidos é necessário que tenhamos redes confiáveis. Assim, é vital a determinação as necessidades dos usuários e a geração dos parâmetros que possam assegurar o atendimento das aplicações dos usuários.

Monitoração e melhor utilização da largura de banda

Resource Reservation Protocol (RSVP)- para que possamos garantir a manutenção da QoS, o RSVP proveêm um conjunto de ferramentas para que na rede os routers trabalhem cooperativamente com mesmo objetivo dos parâmetros de qualidade.

Monitoração e melhor utilização da largura de banda

IPv6 e QoS- A próxima geração de protocolo IP (IPng), ou seja o IPv6, prove em seu datagrama alguns campos que permitem o gerenciamento da largura de banda.

Monitoração e melhor utilização da largura de banda

Problemas de Gerências de Conexão
(*Políticos*)

- Falta de gerência de conexão leva o suporte da rede culpar a prestadora de serviço, equipamentos e enlaces;
- A efetiva gerência com o uso de certa limitação para alguns tipos de tráfego causam problemas para alguns usuários;
- A gerência de conexão é um *jogo de perde e ganha* entre usuários pois a largura de banda não é infinita, não é *free of charge.*

Monitoração e melhor utilização da largura de banda

Problemas de Gerência de Conexão
(Tecnológicos)

Para aplicações de vídeo e áudio é desejável redes onde interrupções sejam mínimas (embora estes serviços possam empregar uma rede como um certo atraso).

Implementações de *redes alto desempenho* devem prover serviços diferenciados para as aplicações diferentes. Um exemplo, são as redes ATM onde a relação de tempo comum para aplicações de áudio e vídeo é estabelecida a uma transferência constante de bits. Numa rede ATM, também, prove serviços de transferência a taxas diferenciadas e conexões orientadas e não-orientadas a conexão.

Monitoração e melhor utilização da largura de banda

Alguns profissionais consideram a gerência de largura de banda rede uma *back art.*
A razão para tal afirmação muitas vezes se baseam no ponto que deve existir uma coordenação global das redes para que o fluxo entre as mesma possam ser parametrizados.

Os oito bits do *type of service* do datagrama IP podem ser usados para prioridade.

Monitoração e melhor utilização da largura de banda

Redes ATM proveêm QoS, nativamente, segundo a recomendação ITU 1350. Esta especificação serve como guia para que fabricantes e usuários empreguem um conjunto de parametros para redes de alto desempenho e para os diferentes tipos de classes de serviços.

Nas redes IP, que não possuem originalmente tal facilidade, os administradores devem gerenciar manualmente os serviços e tipos de tráfego. É convencional a contratação do aumento da largura de banda. Canais dedicados e o uso de protocolos de reserva (ex. RSVP), permitem que os roteadores façam a alocação de banda devida para uma dada aplicação.

Monitoração e melhor utilização da largura de banda

Problema : *O que fazer se o adminstrador da rede não tiver como especificar toda a QoS de aplicação, pois a mesma passa pela Internet ?*

Muito pouco. Dentro da rede o administrador pode empregar as políticas que quiser para atingir uma determinada QoS. Para os limites fora da sua rede e na Internet, mesmo sabendo de todos os parâmetros necessários, seus esforços para obtenção da QoS poderão não ser bem sucedidos. Importante observar, por exemplo, a discrepância entre a largura de sua rede local (10, 100 ou até Gbps) para a enlace WAN (1, 2 ou até centenas de Mbps).

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes com IPv4

A versão 4 do IP possui poucas ferramentas para a atribuição de prioridades para os pacotes IPv4. As ferramentas nativas do Ipv4 são de pouco eficácia e pacotes adicionais são,geralmente, usados para a gerência dos pacotes IP.

Importante observar que as ferramentas existentes (um exemplo é o protocolo RSVP) para a gerência do Ipv4 tem um custo alto e são muitas vezes complexas para redes pequenas. Estas são utilizadas em redes de grande escala que, comumente, apresentam os maiores problemas no tocante a Qualidade de Serviço.

Monitoração e melhor utilização da largura de banda

Resource Reservation Protocol (RSVP)

Este protocolo prove *certo nível* de controle sobre o fluxo de dados. O RSVP é baseado em roteamento (*routed-based*) permitindo aos roteadores fazer solicitações a outros roteadores. Está sob desenvolvimento no IETF.

Para maior eficiência da rede, *todos* os roteadores devem suportar o RSVP.O protocolo quando faz uma solicitação, não existe uma garantia de atendimento, uma vez que a largura de banda já pode estar alocada para outro tráfego da rede que possui uma prioridade maior do que a do solicitante.

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes com IPv4

Resource Reservation Protocol (RSVP)

Os nós recebedores do protocolo RSVP podem fazer solicitações para outros roteadores, entre ele e o outro remetente para o estabelecimento da *reserva do serviço* num sentido do fluxo de tráfego. Os roteadores recebedores de solicitações devem com um certa periodicidade fazer *reserva* para garantir que os roteadores ao longo do caminho estão *cientes* da reserva.

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes com IPv4

Resource Reservation Protocol (RSVP)

As solicitações RSVP são semelhantes aos mecanismos do ICMP usados pelo IP. A medida que uma solicitação vai passando pela rede, roteadores ao longo do caminho indicam quais os serviços eles podem suportar. A solicitação também ajuda na determinação da MTU que será empregada para o fluxo. As informações tratadas são :

- *Token Bucket Rate and Token Bucket Size;*
- *Peak Data Rate and MinimumPoliced Unit;*
- *Maximum Packet Size*

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes com IPv4

Resource Reservation Protocol (RSVP)

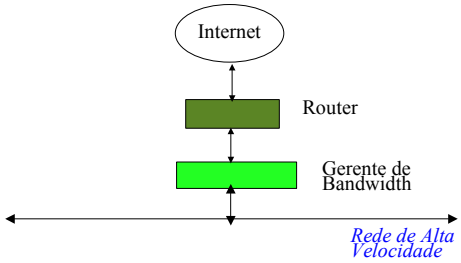
A solicitação RSVP está *submetida* as decisões de roteamento dos protocolos *Open Shortest Path First (OSPF) e Border Gateway Protocol (BGP).* Em outras palavras, primeiro é efetuado

o algoritmo pelo qual deverá ser o caminho do pacote *RSVP* só depois é que os pacotes poderão solicitar a qualidade de serviço necessária.

Desta forma, o RSVP pode ainda não representar um ganho para as redes que usam o IPv4. Mais referências

www.ietf.org/html.charters/*rsvp-charter.html*

Monitoração e melhor utilização da largura de banda



Monitoração e melhor utilização da largura de banda

Produtos *Bandwidth Manager :*

- *Aponet Bandwidth Manager* (www.oponet.com)*;*
- *IPATH Active Traffic Manager* (www.thestructure.com)*;*
- *Packeteer Packetshaper* (www.packeteer.com) *;*
- *Checkpoint Floodgate-1* (www.checkpoint.com) *;*
- *Sun Bandwidth Allocator* (www.usec.sun.com/software/band-allocator) *;*
- *Ukiahsoft Trafficware* (www.ukiahsoft.com).

Monitoração e melhor utilização da largura de banda

Aponet Bandwidth Manager

As principais características do produto são :

- prove canais de largura de banda disponíveis para um IP específico e uma dada porta, também provendo facilidade de monitoração de seu uso;
- o administrador pode atribuir limites de entrada e de saída para determinados limites de endereços(ou grupo de endereços);
- É bastante empregado por ISP para a atribuição dos limites de largura de banda de entrada e saída para seu usuários;

Monitoração e melhor utilização da largura de banda

- O Bandwidth Manager vem em duas versões um de 10 e outro de 100 Mbps. O primeiro para menores quantidades de volume de informação e o outro para maiores;
- O BM confia no controle de fluxo do TCP para a gerência do tráfego, descartando pacotes quando ocorrer uma maior volume de dados.

Monitoração e melhor utilização da largura de banda

IPATH Active Traffic Manager -

As principais características do produto são :

- dispositivo montado em rack de 10 e 100 Mbps;
- em caso de falha os dados passam automaticamente pelo dispositivo;
- a gerência pode ser efetuada por uma aplicação Web-based ou linha de comando;
- o produto permite a identificação de grupos de hosts e subnets para atribuição de alocação de bandwidth para o tráfego de entrada e saída para diversos protocolos;

Monitoração e melhor utilização da largura de banda

- é permitido ao adminstrador auferir tipos de tráfego máximo e mínimo para um determinado critério de largura de banda;
- o produto usa uma aplicação compatível com SNMP para seu gerenciamento;

Monitoração e melhor utilização da largura de banda

Packeteer Packetshaper -

As principais características do produto são :

- dispositivo montado em rack de 384kbps, 10 e 100 Mbps;
- a classificação de tráfego pode ser baseado em URL;
- prove o controle sobre uma variedade de protocolos, tipo TCP/IP, IPX, Appletalk, SNA e outros;
- Emprega o TCP Rate Control ao invés do enfileiramento, evitando a perda de pacotes;

Monitoração e melhor utilização da largura de banda

Packeteer Packetshaper -

- Faz uma estimativa da latência da rede, preve o tráfego que chega, ajusta a janela TCP para compensação;
- Faz um optimização no controle do envio de pacotes ACKs;
- Para o protocolo UDP, cria uma fila para ordenação de pacotes que chegam fora de seqüência;
- Suporta até 2000 conexões TCP e 1000 conexões UDP para os modelos de 10 Mbps e 384 Kbps;

Monitoração e melhor utilização da largura de banda

Packeteer Packetshaper -

- Suporta até 20.000 conexões TCP e 10.000 conexões UDP para o modelo de 100Mbps;
- O pacote oferece integração com o HP OpenView e SNMP

Monitoração e melhor utilização da largura de banda

Checkpoint Floodgate -1

- É um BM que complementa o Firewall-1;
- Da mesma forma que o pacote Firewall, o FloodGate-1 verifica qual as prioridades que os pacotes têm;
- Usando uma GUI, o administrador pode determinar regras de tráfego aplicável para diferente hosts, destinos e tipos de tráfego;
- O adminstrador pode configurar até quatro tipos diferentes de tráfego, assim como permissões e exceções;

Monitoração e melhor utilização da largura de banda

Checkpoint Floodgate -1

- Emprega diferentes ferramentas como GUI - para Windows 95, Windows NT e Solaris;
- Quando o tráfego está lento este continua a passar, todavia quando ocorre um congestionamento existe um filtro pelo peso baseado no peso relativo dos pacotes.

Monitoração e melhor utilização da largura de banda

SUN Bandwidth Allocator

- Prove um sistema de monitoração e gerenciamento mínimo ambientes Sun Solaris;
- Existe uma classificação dos pacotes por classes e assim o tráfego é monitorado e gerenciado;
- O pacote pode ser gerenciado remotamente por computador rodando Java na rede;
- Emprega SNMP para o gerenciamento e estatísticas.

Monitoração e melhor utilização da largura de banda

UKIAHSOFT TRAFFICWARE

Este ambiente prove um serviço diferenciado dos demais - a capacidade do gerenciamento da largura de banda sendo efetuada no computador do usuário (e o mesmo operando) e não apenas empregando o endereçamento IP.

O pacote executa em servidores Windows NT 4.0 e utiliza a informação de *login* criada na início da sessão do Windows para estabelecer uma prioridade de tráfego para a rede.

Monitoração e melhor utilização da largura de banda

UKIAHSOFT TRAFFICWARE

O pacote de software opera semelhante aos demais pacotes. A utilização de uma política na origem (baseada no endereçamento IP, ou autenticação do usuário), e o tipo de tráfego (baseado em portas) são os parâmetros levados em consideração para o tráfego da rede.

O Trafficware usa tanto filas, bem como controle de fluxo para controlar o fluxo TCP/IP. Existe uma GUI para efetuar o gerenciamento e emprega o protocolo SNMP.

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes IPv6

Embora o *Internet Protocol* tenha se expandido de maneira nunca imaginada, o protocolo não recebeu nenhuma mudança expressiva desde a RFC 791 de 1981. Todos estes anos causaram uma certa obsolescência no protocolo. Problemas como endereçamento, dificuldades de roteamento e problema de segurança são alguns dos pontos críticos encontrados no IPv4.

O IETF desenvolveu pelas razões apresentadas o IPv6.

Monitoração e melhor utilização da largura de banda

Ferramentas para Gerência de Bandwidth em Redes IPv6

O IPv6, ao contrário do IPv4, tem um conjunto de ferramentas que auxiliam de uma forma mais eficiente o serviço de qualidade (QoS) na rede.

O cabeçalho do IPv6 dispõe de dois novos campos que auxiliam a gerência da QoS, *Flow Label e Priority.*

Monitoração e melhor utilização da largura de banda

FLOW LABEL

Este campo permite que os pacotes que tenham que ter um tratamento diferenciado sejam assim tratados.

O campo tem tamanho de 20 bits, composto do endereço de origem e IP destino, permitindo que os roteadores mantenham o estado durante o fluxo ao invés de estimar a cada novo pacote.

As aplicações são obrigadas a gerar um *flow label* a cada nova requisição. O reuso do *flow label* é permitido quando um fluxo já está terminado ou foi fechado.

Monitoração e melhor utilização da largura de banda

FLOW LABEL

A utilização de campo *flow label* prove aos roteadores uma maneira fácil de manter as conexões e manter o fluxo de tráfego numa mesma taxa.

Monitoração e melhor utilização da largura de banda

Priority

A utilização do campo *priority* prove aos programas a facilidade de identificar a necessidade de tráfego que os estes necessitam.
O uso efetivo, ou normalização, de como este campo junto com o *flow label* devem plenamente operar ainda estão em discussão.

O campo de 8-bits destinado a *Classe* está no momento a nível de desenvolvimento. Todavia, os 4 bits de prioridade podem nos ajuda a entender o que poderemos ter pela frente.

Monitoração e melhor utilização da largura de banda

IPv6 Priority Values

Valor	Descrição
0	Tráfego sem características
1	Tráfego Filler (fluxo contínuo de informação onde o tempo particularmente não interessa)
2	Transferência de dados sem supervisão (ex. e-mail)
3	Reservado
4	Transferência grande de dados com supervisão (ex. HTTP, FTP e tráfego NFS).

Monitoração e melhor utilização da largura de banda

IPv6 Priority Values

Valor	Descrição
5	Reservado
6	Tráfego Interativo (ex. telnet)
7	Tráfego de Controle da Internet (informação usada por dispositivos que fazem parte da Internet, como roteadores, switches e dispositivos que empregam o SNMP para reportar estados).
8-15	Pacotes em processos que não podem controlar congestionamentos. Pacotes com valor 8 serão descartados antes do de valor 15.

(II) Tecnologias de Redes de GIGABIT

(2.1) Protocolos MAC para Alta Velocidade

Com o aumento significativo da largura de banda dos meios de transmissão, novos desafios na utilização destes ambientes surgiram para as arquiteturas de redes de computadores.

Dentre os vários desafios (exemplos são as *interfaces, protocolos e dispositivos de interconexão*), os protocolos de MAC são uma das grandes preocupações para que possamos atingir com eficiência o compartilhamento do meio físico.

Aos clássicos aplicativos que se utilizavam das redes de comunicação (exemplos: a transferência de arquivos, acesso remoto e comunicação via voz), estão rapidamente sendo agregadas novas aplicações, como transferências em tempo real de informação :

- TV de alta definição;
- transferência de vídeo;
- Vídeo telefonia;
- Aplicações multimídia;

O grande problema que nos deparamos pode ser explicado pela distância entre os requisitos de largura de banda das aplicações e as tecnologias de acesso ao meio.

Interessante notar o histórico de largura de banda das redes de comunicação e sua evolução em bps. Os dados apresentados a seguir ilustram um crescimento de magnitude próximo de segunda ordem até os anos 90.
- Anos 60 : 300 bps;
- Anos 70 : 64Kbps;
- Anos 80 : 1,5 Mbps;
- Anos 90 : 150Mbps.

Com a chegada das redes de multi-giga bps as aplicações podem atingir um grau de complexidade nunca antes imaginado (como transmissão em tempo real de imagem, som, texto, etc).

As redes de fibra ótica são as grande responsáveis por larguras de banda nunca imaginadas, estas podendo atingir Tbps. Existem dois *problemas técnicos*, ainda para serem transpostos quanto a transmissão na rede física de tal quantidade de bps :

• A limitação de potência de transmissão (*power bottleneck*) do nó transmissor, enquanto que os nós receptores requerem um mínimo de energia para recebimento do sinal. Este fenômeno limita o número de nós na rede.

• Com referência as interfaces eletrônicas *(electronic bottleneck),* temos uma limitação de recebimento da ordem de Gbps.

Finalizando, para atingirmos os Tbps é desejável a utilização da *Wavelength Division Multiplexing (WDM).*

LANs/WANs e MAC

As características das LANs e WANs devem ser observadas para que tenhamos implementados eficientes MACs.
Diferenças de taxas de erro e algoritmos de roteamento dos pacotes.

LANs : as taxas de erros estão compreendidas entre 10-8 à 10-11, enquanto que o roteamento usual é o *broadcast*;

WANs : os erros, geralmente, ocorrem com uma freqüência entre 10-5 à 10-7 e diferentes (e complexos) algoritmos de roteamento são empregados.

LANs/WANs e MAC

Um outra característica interessante que diferencia as duas tecnologias é a abordagem de utilização.

Nas LANs os administradores procuram evitar o intenso uso dos nós aos mesmo tempo, devido a característica de acesso ao meio (*broadcast*).

Por outro lado, nas WANs o emprego intensivo significa um melhor uso dos enlaces e recursos disponíveis.

LANs/WANs e MAC

Recentemente as LANs vêem sendo classificadas em quatro categorias :

(1) Baixa e Média Velocidade - entre 10 e 20 Mbps;

(2) Alta Velocidade - entre 50 e 150 Mbps;

(3) *Supercomputer LANs* - entre 800 e 1600 Mbps;

(4) Ultragigabit - redes da ordem de Tbps.

LANs/WANs e MAC

Os protocolos de acesso (usualmente) consideram três dimensões, estas :

- O tempo (síncrono ou assíncrono);
- A topologia (barra, anel, estrela, árvore e multicanal);
- Modo de acesso (aleatório, ordenado ou híbrido).

LANs/WANs e MAC

A topologia *estrela* é caracterizada por um nó central, no qual todos as *estações* estão interconectadas. O nó central pode implementar uma ligação passiva (*hub*), ou ativa (*switched*).

As redes com topologia *estrela* são mais apropriadas para as redes de alta velocidade devido a fácil conexão entre seus pontos finais (ligação ponto-a-ponto) através de fibra ótica.

LANs/WANs e MAC

A topologia *árvore (tree)* é uma generalização das configurações em *estrela*. Uma *árvore* consiste de uma estrutura hierárquica, onde as *estações* são as folhas. As *estações* são ligadas a outras *estações* de outro nível por intermédio de enlaces óticos. Todos os níveis são conectados ao nível *maior* da estrutura.

Existe na atualidade um grande interesse no desenvolvimento de redes com topologia em *árvore*. Entre os motivos podemos visualizar a escalabilidade, modularidade e enlace de alta velocidade.

ESTUDO DE CASO
CASO A

Um exemplo de rede com MAC *Collision Avoidance Star and Tree Network* é a conhecida *Supernet.*

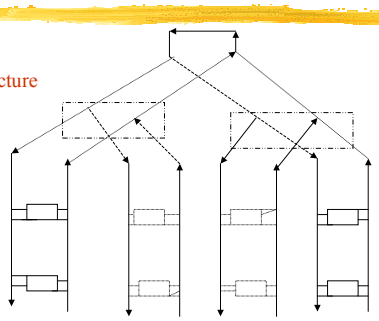
A arquitetura *Supernet* é uma rede a 100 Mbps, consistindo de um par de árvores com a fibra ótica como meio físico de conexão. O par de árvores é conhecido como *selection tree.*

ESTUDO DE CASO
CASO A

Na *Supernet* o roteamento é efetuado por um nó concentrador central. Os nós internos da árvore são denominados de *subhubs* e trabalham de forma análoga ao concentrador central.

As estações são interligadas as duas árvores por intermédio de dispositivos inteligentes chamados de *NAC (Network Access Controllers).*

Supernet Architecture



ESTUDO DE CASO
CASO A

O funcionamento da *Supernet* é o seguinte :

- os pacotes são transmitidos no pedaço de árvore selecionado;
- a transmissão é efetuada através dos *subhubs* até o concentrador principal;
- caso o concentrador principal esteja ocupado, processando outros pacotes, pacotes que chegam são descartados;
- quando os pacotes chegam ao concentrador principal com sucesso os mesmos são propagados através de um *broadcast* no outro lado da árvore;

ESTUDO DE CASO
CASO A

- o nó destino reconhece seu endereço é avisa o recebimento dos pacotes ao *central hub* que repassa para o nó remetente;
- quando ocorre um *estouro de tempo* no *round-trip-time*, o nó remetente assume que os pacotes não chegaram ao destino e transmite os pacotes perdidos.

ESTUDO DE CASO
CASO A

Embora esta arquitetura não seja uma solução ideal, existem vários relatos que indicam que (mesmo para uma carga grande do ambiente) a rede tem um largura de banda sustentada da ordem de 100 Mbps.

ESTUDO DE CASO
CASO A

Uma melhoria da *Supernet* é a implementação conhecida por *Collision Avoidance Multiple Broadcast (CAMB)*. Nesta abordagem os *subhubs* podem atuar como se fosse o concentrador central.

Se um *subhub* reconhece que é um *ancestral* do nó destino, este realiza o trabalho do *root hub*. Em outras palavras, o *subhub* faz o roteamento dos pacotes necessário para atingir o nó destinatário. Caso os pacotes não sejam reconhecidos pelo *subhub* como estando no seu domínio, estes são repassados para o *hub* imediatamente acima.

ESTUDO DE CASO
CASO A

A arquitetura CAMB prove uma melhora na comunicação geral da rede, pois a existência de *clusters* de famílias é muito comum. Em outras palavras, a *concorrência* de comunicação prove um meio eficiente de aproveitamento da rede.

ESTUDO DE CASO
CASO B

Em redes com arquitetura *mesh,* todos as estações estão interligadas por enlaces ponto-a-ponto dedicados. Cada estação possui vários enlaces simultâneos de entrada e saída. Desta forma, um pacote pode ser enviado/recebido através de mais de um canal.

ESTUDO DE CASO
CASO B

As redes em *mesh* são caracterizadas por :
- a possibilidade de *flooding,* quando um (ou mais) pacote (s) são enviados simultaneamente para diversas portas;
- o roteamento pode ser efetuado a cada *nó* dependendo do pacote e do algoritmo de roteamento;
- alto valor agregado de largura de banda devido a seus múltiplos caminhos;
- excelente nível de tolerância a falha.

ESTUDO DE CASO
CASO B

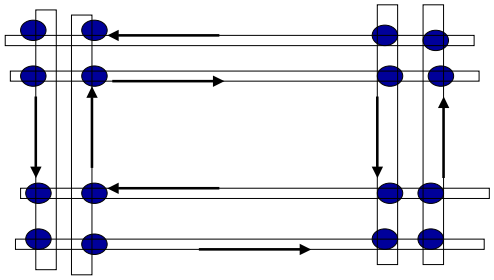
Um exemplo de implementação em *mesh* é o ambiente conhecido por *Manhattan Street Network (MSN).* O MSN consiste de uma estrutura regular com um número de par de linhas e colunas de enlace e as estações estão na intersecção.

ESTUDO DE CASO
CASO B

As estações na mesma linha, ou coluna, são conectadas por *loops* unidirecionais com direções alternadas nas adjacentes linhas e colunas. Assim, cada estação possui dois enlaces de entrada e dois de saída.

Diversos MACs podem ser empregados no ambiente MSN, entre os mais comuns estão o *slotted e token-passing.*

Manhattan Street Network

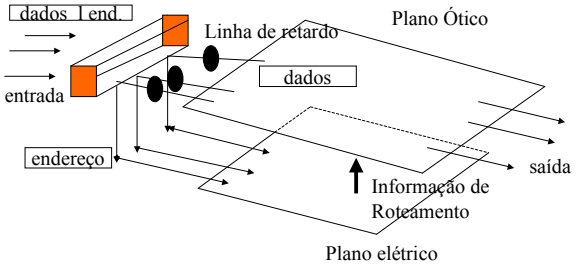


(2.2) Arquitetura de Redes Óticas

As redes que se utilizam de fibra óticas, conhecidas de uma forma genérica como *optical network architectures*, são o estado da arte das redes de alto-desempenho.

Todavia, estas redes sofrem de um *mau congênito - os gargalos óticos eletrônicos.* Este mau inibe o uso pleno da tecnologia de fibra ótica para o transporte de Tbps.

Nas redes óticas a informação é *modula em luz*. Numa rede completamente ótica(*all-optical network*) o sinal de luz é empregado para todas as operações. Desta forma não apenas o *payload,* mas também as operações de controle como o roteamento, comutação e manutenção são efetuadas empregando a luz.



Exemplo de Arquiteturas Óticas

A seguir apresentamos alguns exemplos de arquiteturas óticas e suas respectivas funcionalidades :

• Redes *Circuit-Switched* - redes óticas com reconfiguração lentas podem ser empregadas para aplicações de suporte para vídeo e gráficos. Exemplos são as redes IBM *Rainbow, DiSCO da AT&T.*

• Redes *Packet-Switched* - redes que necessitam de comutação da *ordem* de centenas de nsec devem ser utilizadas, pois a rápida reconfiguração não *pesa* em relação ao tamanho do pacote. Implementações conhecidas são HYPASS e a LAMBDANET da Bellcore.

Exemplo de Arquiteturas Óticas

A seguir apresentamos alguns exemplos de arquiteturas óticas e suas respectivas funcionalidades :

• Redes *Circuit-Switched* - redes óticas com reconfiguração lentas podem ser empregadas para aplicações de suporte para vídeo e gráficos. Exemplos são as redes IBM *Rainbow, DiSCO da AT&T.*

• Redes *Packet-Switched* - redes que necessitam de comutação da *ordem* de centenas de nsec devem ser utilizadas, pois a rápida reconfiguração não *pesa* em relação ao tamanho do pacote.

Fiber Channel

É um padrão da industria empregado para a transmissão como rede de interconexão, rede local e suporte a transmissão tradicional de E/S. Como o nome diz o padrão foi originalmente projetado para usar fibra, embora a especificação inclua cabo coaxial e par trançado para pequenas distâncias.

A largura de banda compreende uma faixa que vai de 12.5 à 100 Mbytes por segundo. A distância máxima sem repetidores é de 10 Km.

O padrão FC suporta protocolos de mais alto nível, tais como :

(1) SCSI;

(2) HIPPI-FR (Framing Protocol) ;

(3) IP;

(4) IBM System/390 I/O.

Protocolo de Alto Nível	
FC - 4	Mapeamento
FC - 3	Serviços Comuns
FC - 2	Sinais Lógicos
FC - 1	Transmissão
FC - 0	Físico

O FC é dividido em cinco níveis funcionais, denominados de FC - 0 até o FC - 4 e têm as seguintes funções :

- FC - 0 : define o nível físico - cabos, conectores, velocidade de bits, especificação ótica e elétrica, especificação de jitter e distância sem repetição;
- FC - 1 : neste nível é definido a codificação dos dados e controle, sincronização de bit/byte e palavras, controle de erros;
- FC - 2: define a forma do protocolo de sinalização, é semelhante ao enlace nos protocolos convencionais. Todavia com o atual hardware o enlace é efetuado ponto-a-ponto;

- FC - 3 : define o nível de serviço entre nodes;
- FC - 4 : faz a interface entre os protocolos de mais alto nível e as camadas FC - 2 e FC 3. Um exemplo é a transmissão de um pacote IP, ou seja esta camada auxilia a transmissão utilizando os serviços da camada FC - 2 .

O HIPPI é um padrão que visa a transferência direta entre memoria com taxas de transferência entre 800 Mbps e 1600 Mbps

HIPPI - LE (link Encapsulation)	Outros LLC <> 802.2
HIPPI - FR (Framing Protocol) Formato do pacote e Header	
HIPPI-PH (elétrico, mecânico e sinalização)	HIPPI -SC (Switch Control) Nível Físico de C.S.
Serial - HIPPI Fiber-based HIPPI-PH extendida a 10 Km	

Proposta HIPPI

Para melhor compreensão da tecnologia VIA, esta apresentação está baseada no trabalho *Evolution of the* Virtual *Interface Architecture,* de Thorsten von Eicken and Werner Vogels, Cornell University, IEEE Computer - November. 1998

Cluster

- Webster´s New World - Dicionário de Informática - 6ª Edição, 1998: Em um disquete ou disco rígido, a unidade básica de armazenamento de dados. Um cluster abrange dois ou mais setores.
- Webster´s New World - College Dictionary - Third Edition, 1997: A number of things of the same sort gathered together or growing together.

Cluster

- VIA Architecture Home Page, 1997:
  - Cluster: the linking together of individual servers and workstations into a cohesive unit.
  - SAN (high-performance system area network): a specialized network optimized for the reliability and performance requirements of clusters.

Introdução

- Motivações
- VIA Organization
  - Promotores
    - ***Compaq Computer Corp.***
    - ***Intel Corporation***
    - ***Microsoft Corp.***
  - Contribuintes
    - Mais de 100 organizações e indústrias
- Disponível desde 16 de dezembro de 1997

Introdução

- Influência de diversos protótipos - entre os quais destaca-se U-Net (Cornell University).
- Interface de rede GNN1000, da GigaNet (http://www.giganet.com).

Introdução

- Visão Geral
  - Como fornecer às aplicações acesso direto à interface de hardware da rede, mantendo proteção suficiente para garantir que as aplicações não interfiram umas nas outras.

Introdução

- Como projetar uma interface de programação eficiente e, ainda assim, versátil ?
  - As aplicações devem ser capazes de acessar a interface de rede e ainda assim controlar o *buffer*, o escalonamento e o endereçamento e
  - A interface de programação deve suportar uma grande variedade de implementações de hardware.

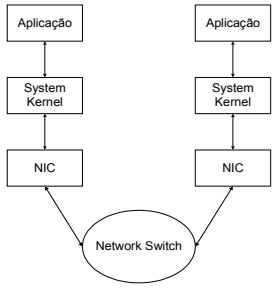
Introdução

- Visão Geral
  - Como gerenciar os recursos, em particular a memória - custos de transferências da DMA para mapeamento de endereços virtuais em físicos.
  - Como gerenciar o acesso imparcial à rede sem o uso do *kernel*, que, na pilha de protocolos tradicional, atua como ponto de controle central e escalonamento.

Introdução

- Preparando o ambiente
  - Remover o *kernel* e sua pilha centralizada de rede do caminho crítico.
  - Criar uma interface de rede em nível de usuário (*user-level network interface*)
    - Diminui a latência na comunicação e
    - Aumenta o *throughput* na rede.

Modelo Tradicional de Redes



Fatores de Desempenho

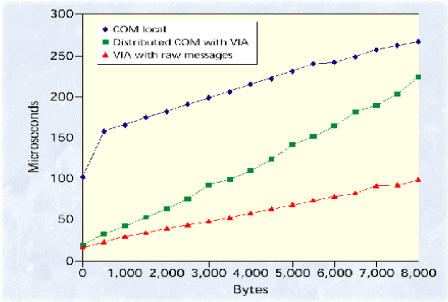
- O que é Desempenho?
  - Largura de banda alcançada quando um fluxo infinito está sendo transmitido (definição tradicional).
- Fatores importantes para uma correta definição
  - Latência na comunicação;
  - Largura de banda para mensagens pequenas e
  - Utilização de um protocolo flexível.

Fatores de Desempenho

- Baixa latência na comunicação
  - É a chave para utilização de clusters.
  - Causada principalmente por *overhead* dos processos:
    - gerenciamento de *buffer*;
    - cópia da mensagem;
    - computação de *checksums*;
    - manipulação do controle de fluxo e interrupções e
    - controle das interfaces de rede.

Round Trip Time

Communication Latency



Fatores de Desempenho

Elevada largura de banda para mensagens pequenas
- Mensagens pequenas ≈ 1 Kbyte.
- Ao reduzir o *overhead* das mensagens, as interfaces de rede em nível de usuário podem fornecer toda a largura de banda para as menores mensagens possíveis.

Fatores de Desempenho

Protocolos de comunicação flexíveis
- Dois pontos fundamentais:
  - a integração da aplicação e do gerenciamento de *buffer* do protocolo e
  - a otimização do protocolo de controle do caminho.

Fatores de Desempenho

Protocolos de comunicação flexíveis
- Como obter um protocolo flexível?
  - Usar versões experimentais ou altamente otimizadas de protocolos tradicionais;

Fatores de Desempenho

Usar técnicas avançadas para projetar o protocolo:
- no nível de aplicação de quadros (*frames*);
- na camada de integração de processamento;
- compilação de todos os protocolos conjuntamente, ao invés de um pequeno subconjunto de protocolos específicos e
- otimização da geração do *path*.

Fatores de Desempenho

- Distância entre CPU e NIC
  - A NIC possui um barramento de I/O mais lento.
  - Tempo necessário para atravessar os barramentos.
  - Alternativas:
    - Colocar a NIC em um barramento mais rápido e
    - ON-CHIP NIC.

Trade-Offs

- São mais perceptíveis em:
  - Estrutura de filas
  - Estratégia de gerência de memória
  - Multiplexação e Demultiplexação
  - Acesso remoto à memória (RDMA)

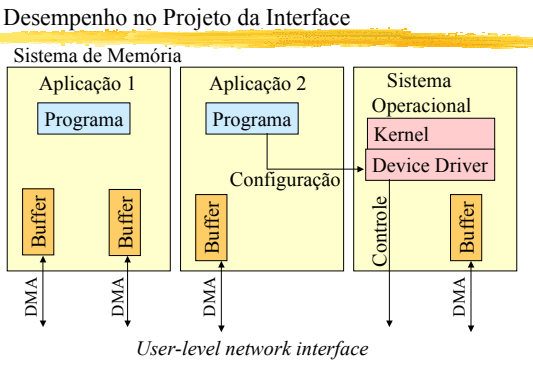
Desempenho no Projeto da Interface

- *User-level network interface*:
  - está baseada em mensagens; realiza a troca dados recebendo-os e enviando-os de forma explícita, similar ao modelo de troca de mensagens tradicional;
  - permite que múltiplos do *host* acessem simultaneamente a rede;

Desempenho no Projeto da Interface

A interface de rede em nível de usuário:
por motivos de segurança, separou-se a configuração da comunicação da transferência dos dados. Assim:
configuração: o sistema operacional é invocado para garantir que uma aplicação não possa interferir na outra;
durante a transferência dos dados, a interface desvia o sistema operacional e executa uma verificação simples para forçar a segurança.

Desempenho no Projeto da Interface



Desempenho no Projeto da Interface

Origem da computação paralela
A *user-level network interface* faz a mediação entre a rede e a aplicação:
como a aplicação especifica a localização da mensagem a ser enviada
onde estão alocados buffers livres para recepção
como a interface notifica à aplicação que chegou uma mensagem

Desempenho no Projeto da Interface

Origem da computação paralela
A *user-level network interface* faz essa mediação de duas formas:
send / receive: *Active Message* e *Fast Message*
filas de processos: U-NET e VIA

Desempenho no Projeto da Interface

Origem da computação paralela
- o modelo de troca de mensagens utilizado na computação paralela tem sua origem no modelo tradicional
- para enviar um dado é preciso especificar:
  - o endereço de memória original do dado
  - o nodo processador destinatário

Desempenho no Projeto da Interface

Origem da computação paralela
- para receber um dado é preciso:
  - transferir explicitamente a mensagem que chega para uma região de memória do destinatária.
- Por causa da semântica das operações de transferência, a *user-level network interface* deve:
  - *bufferizar* as mensagens ou
  - executar um *handshake* entre o emissor e o destinatário da mensagem.

Desempenho no Projeto da Interface

Origem da computação paralela
- As *Active Messages* foram criadas para endereçar o *overhead* causado pela semântica das operações *send* e *receive*
- idéia principal: colocar um endereço de um *handler* dedicado dentro de cada mensagem e ter a interface de rede chamando o *handler* assim que a mensagem for recebida.

Desempenho no Projeto da Interface

Origem da computação paralela

Assim, *Active Messages* não precisam:
- fazer a bufferização da mensagem e pode confiar no *handler* para busca-las da rede
- preocupar-se com o controle de fluxo ou retransmissão de mensagem, porque as redes, nas máquinas paralelas já implementam esses mecanismos.

Desempenho no Projeto da Interface

Origem da computação paralela
- As *Fast Messages* resolvem o problema de overhead das *Active Messages* substituindo o *handler dispatch* com *buffering* e um *poll* explícito de operações.
- Ambas fornecem operações *send* e *receive* através de chamada de função.

Desempenho no Projeto da Interface

U-NET
- Fornece uma interface para a rede que é próxima às funcionalidades tipicamente encontradas nas interfaces de *hardware* da LAN.
- A U-NET não
  - aloca buffers;
  - executa qualquer mensagem implícita de buffer;
  - despacha qualquer mensagem

Desempenho no Projeto da Interface

U-NET

Consiste de um conjunto de filas na memória que transporta mensagens para a rede

- Diferenças:
  - *Active Message* e *Fast Message* definem um camada fina de software;
  - U-NET especifica a operação do hardware.

Desempenho no Projeto da Interface

U-NET

- Proteção
  - Não é permitido enviar mensagens com destino arbitrário; nem
    receber mensagens endereçadas a outro nó
  - o processo deve iniciar o canal de comunicação antes de enviar ou receber mensagens

Desempenho no Projeto da Interface

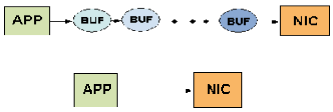
Memória Compartilhada

- Dois modelos fornecem as primitivas de comunicação baseados em memória compartilhada:
  - AM-II (*Active Message* II)
    desenvolvido pela Universidade da Califórnia
  - VMMC (*Virtual Memory Mapped Communication*)
    desenvolvido pela Universidade de Princeton, como parte do projeto de *Cluster Shrimp*

Desempenho no Projeto da Interface

Memória Compartilhada

- As primitivas eliminam a cópia que a *Fast Message* e a U-NET precisam ao final do recebimento.



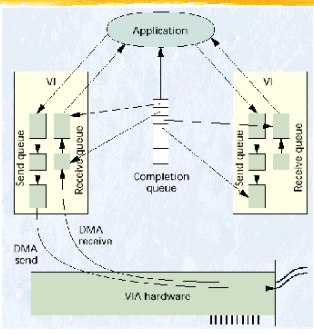
Desempenho no Projeto da Interface

Virtual Interface Architecture - VIA

- combina as operações da U-NET
- Os processos abrem uma VI que representa os handl da rede.
- Cada VI representa uma conexão para uma única VI remota.

Desempenho no Projeto da Interface

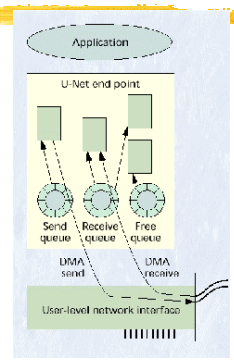
VIA
- Send List
- Receive List



Desempenho no Projeto da Interface

U-Net

- Send Ring
- Receive Ring
- Free Ring



Trade-Offs

Estrutura de Filas

Nesta estrutura uma camada de software mais alta pode gerenciar os descritores e *buffers* requeridos diretamente.
Permite que a transmissão e recepção ocorram assincronamente.
São projetadas para aceitar descritores *scatter-gather,* que permitem às aplicações construírem mensagens a partir de uma coleção de *buffers* não contíguos.

Trade-Offs

Estratégia de gerência de memória

Integrar o gerenciamento de *buffer* entre a aplicação e a interface de rede é um fator importante na eliminação de cópias de dados e na redução de alocações e desalocações.
O gerenciamento de memória da VIA é inspirado na UTLB do VMMC-2, que torna a aplicação responsável por gerenciar a tradução dos endereços virtuais para físicos.

Trade-Offs

Multiplexação e Demultiplexação

Na *user-level network interface*, a NIC deve demultiplexar as mensagens que chegam na fila de recebimento correta.
Quando enviando uma mensagem, a interface deve garantir a proteção.
A multiplexação da VIA é orientada a conexão. Duas VI precisam estabelecer uma conexão antes que qualquer dado seja transmitido.
A U-NET não é orientada a conexão.

Trade-Offs

Acesso Remoto à Memória (RDMA)

Adiciona uma complexidade considerável às implementações VIA. Quando recebendo um *write*, a interface deve não apenas determinar a destinação correta, mas também extrair o endereço de memória destino da mensagem e traduzi-lo.
A interface também pode receber pedidos de *read* em uma taxa maior do que consegue servir. Assim, a VIA deverá ser restrita a implementações que forneçam entrega confiável.

Conclusões

A VIA foi desenvolvida a partir das principais pesquisas sobre interfaces em nível de usuário e fornece uma especificação coerente.

Problemas

Não oferece interoperabilidade com redes que usam interfaces de rede convencionais.
UTLBs são fáceis de implementar, mas retiram recursos que seriam utilizados em paginação genérica.

(2.6) Myrinet

Myrinet é uma System Area Network (SAN), que pode ser entendida como uma rede de interconexão a 1.28 Gbps, full-duplex, proprietária da empresa Myricom (www.myri.com).

Esta SAN utiliza roteamento *wormhole*(ou *cut-through switching*), com uma latência muito baixa, tolerante a falhas com o mapeamento automático da configuração da rede.O ambiente suporta Linux e Windows NT, Irix, DEC Unix, TCP/IP, MPICH, Active Messages.

O ambiente Myrinet tem seu preço mais alto, quando comparado com uma rede Fast Ethernet, o custo por nó é estimado em US$ 1,500. Os switches Myrinet têm no máximo 16 portas, embora possam ser interligados para a construção de um cluster.

Em contrapartida, temos as seguintes características :

- a latência é da ordem de 5 µs numa direção ponto-a-ponto ;
- para maior flexibilidade existe um processador programável na placa;
- pode aproveitar o máximo de um barramento PCI.

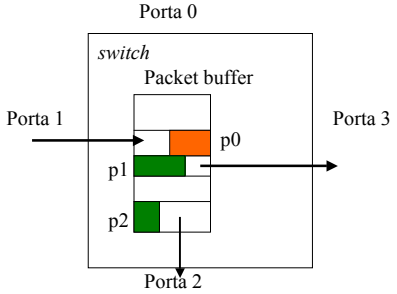
Técnicas de Switching

A conexão entre portas de entrada e saída e a transferência de dados entre estas dentro de uma ambiente de rede é denominado de switching. Existem atualmente duas técnicas de switching, estas a *packet* e a *wormhole switching*.

Técnicas de Switching

Packet switching é o método no qual o pacote completo é armazenado na rede comutada antes que os dados sejam enviados para outro estágio da rede. O mecanismo de store/forward requer um limite máximo de tamanho de pacote (MTU) e algum espaço de buffer para armazenar um ou mais pacotes temporariamente. Redes (LAN/WAN) mais tradicionais empregam este modelo (Fast Ethernet e ATM)

Técnicas de Switching - Packet Switching



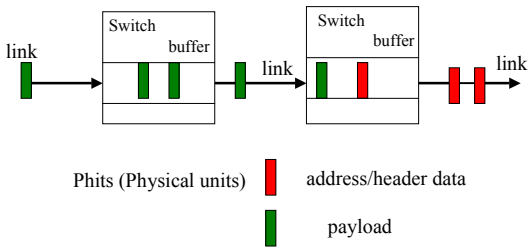
Técnicas de Switching

Wormhole Switching (também conhecido como *cut-through*) é caraterizado pelo envio imediato do dado para o próximo estágio assim que endereço de dados é decodificado. Nesta técnica a baixa latência e a pequena quantidade de buffer são as principais vantagens.

Técnicas de Switching

No exemplo a seguir, ilustra os dados de uma mensagem espalhados por muitos links e switches. Os Phits são a unidade de transmissão no ambiente.O tamanho da mensagem pode ser variável, mas o projetista deve lembrar que mensagens longas pode ser bloqueadas ao longo de vários estágios e, também, a correção de erros é mais difícil de ser trabalhada.

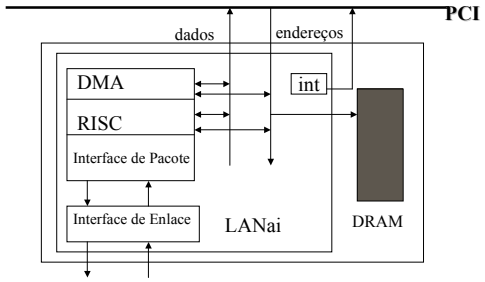
Técnicas de Switching - Wormhole Switching



Técnicas de Switching

Na técnica de packet switching pode verificar o pacote inteiro antes de enviar para um outro estágio. Por outro lado, no caso do *wormhole switching* dados corrompidos serão re-encaminhados erroneamente.

A Interface Host Myrinet



A interface de host Myrinet consiste de duas parte : o chip LANai e a memória SRAM.

O LANai é um chip VLSI customizado que controla a transferência de dados entre um host e a rede. O componente principal é um microprocessador RISC que controla a operação de DMA:
• entre a interface e memória local;
• e o host e a rede.

Qualquer mensagem é primeiro escrita na memória da placa para só depois ser colocada na rede.

(III) Protocolos de Alto-Desempenho

(3.1) Protocolos *Lightweight para Rede de Alto Desempenho*

A prática tem demonstrado que não existe uma correlação direta entre a melhoria da infra-estrutura das redes é um melhor desempenho das aplicações.

Os protocolos têm sido o grande vilão deste problema.Os pacotes de protocolos existentes são inadequados para evolução para ambientes de alto desempenho.

As três abordagem que seguem podem melhorar o desempenho de uma determina arquitetura de protocolo:

- Minimizar os requerimentos de processamento para redução dos custos de transmissão;
- Diminuir o controle de erros para redes que são consideradas sem erro (*error free*) ;
- Melhorar o algoritmo de controle de fluxo.

Novos protocolos denominados de *lightweight protocols,* ou protocolos de alto desempenho, têm sido desenvolvido visando :

- Melhoria na gerência da conexão;
- Controle de fluxo;
- Tratamento de erros.

Melhoria na Gerência da Conexão

Os procedimentos de sinais necessários para a abertura, manutenção e fechamento de conexão entre entidades de comunicação.

Controle de Fluxo

Para evitar congestionamento, um mecanismo eficiente de controle de fluxo é necessário para solucionar a equação com as incógnitas :

- X - Quanto a aplicação no remetente pode solicita de *bandwidth* ?
- Y - Quanto a rede pode dispor de *bandwidth* ?
- Z - Quanto no destinatário a aplicação pode ser processada ?

Tratamento de Erro

Os esquemas existentes para o tratamento da perda de dados na transmissão e *buffer overflow,* devem ser eficientes não só na detecção, mas também na sua recuperação sem causar *overhead* na rede.

Em outras palavras é desejável que um algoritmo eficaz seja empregado para evitar a perda de largura de banda para o controle de tratamento de erro.

Protocolos de Transporte

Devido ao padrão *de facto* do protocolo de rede IP e seu *peso* em todos os ambientes inter-redes, muito se tem desenvolvido ***a nível de transporte***.

A idéia é que eu possa melhorar a utilização da largura de banda através das conexões eficientes de transporte. Assim não só novos protocolos têm sido propostos, mas também há um grande esforço na melhoria daqueles já existentes.

Protocolos de Transporte

Vamos estudar oito protocolos de transporte devido a sua importância e interessantes diferenças quanto a sua adequação como um pacote de alto desempenho.

Protocolos de Transporte

As arquiteturas de protocolos estudadas serão :

•APPN;
• Datakit;
• Delta-t;
• NETBLT;
• OSI/TP4;
• TCP;
• VMTP; e
• XTP.

APPN

O Advanced Peer-to-Peer Networking (APPN) é um protocolo de transporte da IBM para os sistemas S/36 e AS/400 construído para integrar a arquitetura de conexão fim-a-fim System Area Network (SNA). As funções de transporte foram implementadas num serviço orientado a conexão virtual baseado na total confiabilidade da conexão de enlace.

APPN

O APPN não dispõe de serviço de tratamento de mensagem, assim é um bom exemplo de protocolo que *confia* nos serviços de alta qualidade das camadas inferiores da rede.

Datakit

O protocolo foi desenvolvido como sendo um *protocolo de transporte universal*, independente de uma aplicação específica ou ambiente.
O protocolo é baseado num serviço orientado a circuito virtual que entrega os pacotes sem erros e em seqüência, considerando uma possível perda.
A chave deste protocolo são as funções de controle fim-a-fim de controle de fluxo, detecção e retransmissão de dados perdidos.

Delta-t

O protocolo foi desenvolvido para prover uma comunicação eficiente no Lawrence Livermore Labs para serviços ponto-a-ponto confiável, orientados a grande quantidade de dados em redes não orientas a conexão.

Delta-t

A inovação deste protocolo foi o desenvolvimento um sistema de gerenciamento de conexão baseado em tempo, através do qual o *Delta-t* suporta conexões *lightweight* com um mínimo de demora de sinalização.

NETBLT

O *NETwork BLock Transfer Protocol* (NETBLT) foi desenvolvido para o transferência de grandes quantidades de dados. Da mesma, o protocolo pode operar com eficiência em redes com grande latência, como redes de satélites empregando o IP.

A conexão NETBLT é unidirecional e é normalmente fechada pelo remetente. A unidade de transmissão é um largo *buffer*. A concorrência de diversas unidades de transmissão é possível o que mantêm o fluxo num nível aceitável.

NETBLT

O controle de fluxo é efetuado na janela de transmissão através de parâmetros de tempo iniciais no intervalo de negociação.

O tratamento de erros emprega uma retransmissão seletiva. Uma única solicitação de retransmissão pode ativar um número arbitrário de pacotes para serem retransmitidos.
Ao final de uma transmissão, todos os *NACK* são enviados como um pacote único.

OSI/TP4

O TP4 é um protocolo desenvolvido sob coordenação da OSI, no qual as seguintes características interessantes existem :

- grande número de parâmetros de QoS : entre estes existem throughput, percentual de erro, prioridade e demora de transmissão;
- uma conexão de transporte pode ser dividida em várias conexões, ou seja multiplexando a saída das solicitações.

Datakit

O protocolo observando que um dos pontos de gargalo numa comunicação são os *recebedores* no seu processamento dos protocolos, o *Datakit* utiliza um protocolo denominado *Universal Receiver Protocol (URP).*

Na abordagem URP, somente são respondidos os comandos do transmissor que foram emitidos segundo os serviços oferecidos pelo protocolo de transporte. A unidade de transmissão é o byte, o nono bit é empregado para distinguir entre dados e sinal de controle.

TCP

O IP prove um serviço unificado de datagrama, independente das diferentes *subnets*.A idéia do desenvolvimento de protocolo não orientado a conexão e independente de uma rede particular fez com que roteadores com preços relativamente baixos pudesse rapidamente estar disponíveis.

Desta forma, o protocolo de transporte TCP foi implementado para assegurar uma ligação fim-a-fim confiável orientada a conexão.

VMTP

O *Versatile Message Transfer Protocol* foi desenvolvido para prover a infra-estrutura de comunicação para um dado sistema operacional distribuído. Em outras palavras, o foco principal do protocolo é o suporte conexões orientadas a transações (ex. RPC). Estas aplicações requerem respostas rápidas para pequenas quantidades de dados.

Por outro lado, o VMTP também oferece serviço para transferência grande de dados. O controle de erro é seletivo e ainda tem uma facilidade de controle de transferência. O protocolo ainda prove um serviço de *multicast*.

XTP

O Xpress Transport Protocol (XTP) foi primeiramente desenvolvido com propósito para implementações VLSI.O protocolo foi projetado para atender com eficiência um grande espectro de serviços, tais como:

• datagramas em tempo-real;
• multicasting;
• transferência de grande quantidade de informações.

XTP

O XTP oferece a nível de controle :

• controle de fluxo;
• retransmissão seletiva;
• estabelecimento de conexão explicita;

Análise das Funções do Protocolo de Transporte

Nesta seção vamos examinar com mais detalhes os principais mecanismos do protocolo de transporte. Dentre estas funções :

- Gerência de conexão : estudo de início e término de uma associação de transporte;

- Fase de transferência de dados : recebimento de reconhecimento, controle de fluxo e tratamento do erro;

Gerência de Conexão

Visando compreender de uma maneira mais precisa como é efetivamente realizada a gerência de conexão, vamos estudar os seguintes aspectos :

• Sinalização
• Configuração inicial e fechamento
• Seleção do serviço de transporte
• Multiplexação
• Controle da informação
• Formato do pacote
• Troca mínima de pacote

Gerência de Conexão

Sinalização

A troca de informação entre duas entidades de transporte com o propósito de gerência de conexão é conhecida como *sinalização*.

A sinalização pode ser efetuada de maneira que dentro de uma mesma associação, tenhamos dados e informação.A este tipo de abordagem denominamos de *in-band*.

Por outro lado, numa associação *out-of-band* temos os dados e a informação de controle transmitidos separadamente.

Gerência de Conexão

Sinalização

Considerando os protocolos de alto desempenho que foram apresentados e a taxonomia de sinalização, podemos fazer a seguinte consideração :

• *In-band* : TCP, NETBLT, XTP, OSI/TP4, Delta-t, VMTP ;
• *Out-of-band* : Datakit, APPN.

O protocolo VMTP é híbrido quando a seu funcionamento, ou seja efetua a verificação de conexão da forma *out-of-band*. Todavia, o estabelecimento da conexão é *in-band*.

Gerência de Conexão

Sinalização

A conseqüência maior da sinalização *in-band* é que as entidades responsáveis pela conexão têm que resolver a cada pacote se existe, ou não, informação de controle. Este fato acarreta num aumento do processamento normal dos pacotes de dados , fato que não é desejável numa rede de alto desempenho.

Desta uma forma oposta, na sinalização *out-of-band* temos uma desvinculação de dados e informação de controle, ocasionando num fator diferencial para redes que podem multiplexar a níveis mais baixos os pacotes.

Gerência de Conexão

Sinalização

No caso da abordagem *out-of-band* é ainda permitido que diferentes tipos de dados sejam suportados na conexão.

Um fator que aumenta a importância desta facilidade é a necessidade comercial de algumas aplicações de cobrança e segurança de uma única vez (*não se esqueçam como vocês compram livros pela Internet*).

Gerência de Conexão

Sinalização

A sinalização *out-of-band* parece ser a opção correta para as redes de alto desempenho uma vez que o tempo de processamento de pacotes é reduzido.

Gerência de Conexão

Configuração Inicial e Fechamento

Em um estabelecimento de conexão, ou configuração inicial, e seu fechamento estão condicionados a dois mecanismos :

• Handshake - procedimento que requer explicitamente a troca de mensagens entre as entidades de comunicação;

• *Implicit (ou Timer-based)* - neste tipo de esquema a abertura de conexão é efetuada ao primeiro pacote recebido. O fechamento é consumado por intermédio de controle de tempo.

Gerência de Conexão

Configuração Inicial e Fechamento

Importante de observar que no caso da conexão implícita , apenas no caso da abertura com garantia de controle de tempo o esquema funciona de modo apropriado.

Em outras palavras, é necessário que pacotes que tenham um atraso não sejam confundidos como um fechamento. É necessário que a rede *conheça* os atrasos que possam estar ocorrendo.

Gerência de Conexão

Configuração Inicial e Fechamento

Os protocolos podem ser assim classificados :

• *Handshake*
- three-way : TCP, OSI/TP4 (setup), XTP(release)
- two-way : NETBLT, Datakit, APPN, OSI/TP4 (release)

• *Implicit :* Delta-t, VMTP, XTP (setup)

Gerência de Conexão

Configuração Inicial e Fechamento

Considerações (*vamos discutir*) :

- Para aplicações com granulosidade grossa o *handshake não é* tão significante;
- O handshake pode ser atingido *out-of-band*, no caso do *implicit* este é efetuando *in-band* quando o primeiro pacote chega;

Gerência de Conexão

Seleção de Serviço de Transporte

O serviço do transporte tem a responsabilidade na escolha *do que* prover para a aplicação dado uma determinada infra-estrutura de rede.

Dependendo da rede os serviços podem variar de maneira sensível.

Gerência de Conexão

Seleção de Serviço de Transporte

Os seguintes parâmetros devem ser considerados :

- tamanho máximo de pacote;
- valores de *timeout;*
- contadores de tentativas;
- tamanho de buffers.

Uma vez negociados estes parâmetros podem ficar estaticamente estabelecidos durante a conexão.

Gerência de Conexão

Seleção de Serviço de Transporte

Outros parâmetros tais como :

- fluxo de controle;
- número de seqüência*;*

Devem ser continuamente atualizados durante a transferência de dados, pelos algoritmos de controle de fluxo e sistema de recebimento de mensagem.

Gerência de Conexão

Seleção de Serviço de Transporte

Os seguintes parâmetros são considerados pelos protocolos :

- Parâmetros de negociação durante o estabelecimento de conexão : APPN, Datakit, NETBLT, OSI/TP4, TCP, VMTP e XTP;
- Atualização dos parâmetros durante a transferência de dados : Datakit, Delta-t, NETBLT, OSI/TP4, TCP, VMTP e XTP;
- Seleção dos modos de operação : Datakit, VMTP e XTP.

Gerência de Conexão

Multiplexação

A multiplexação é a combinação dos dados de mais de uma conexão a nível de protocolo para uma simples associação.

A multiplexação é efetuada durante a fase de transferência de dados, todavia a facilidade é efetuada durante a fase de estabelecimento da conexão.

Gerência de Conexão

Multiplexação

Os protocolos podem ser classificados segundo a sua multiplexão. A Multiplexação nas conexões na camada de transporte para um ponto da camada de rede é considerado um circuito virtual para as redes orientadas a conexão.

No caso de uma rede não orientada a conexão, significa um par (endereço de origem e endereço destino).

Gerência de Conexão

Multiplexação

Classificação dos protocolos :

• Fazem multiplexação : XTP, Delta-T, VMTP, OSI/TP4, TCP e NETBLT;

• Não Fazem : APPN, Datakit.

Gerência de Conexão

Controle da Informação

O controle da informação é usado para o efetivo sincronismo de estado entre remetente e destinatário. Este serviço é vital para que os controle de gerenciamento de conexão, recebimento de dados, reconhecimento na chegada de pacotes, fluxo de transmissão e tratamento de erro.

Gerência de Conexão

Controle da Informação

Exemplos são :

• O XTP permite que destinatários apontem para lacunas nos dados recebidos. Desta forma, é possível a utilização do algoritmo seletivo de retransmissão.

• O uso excessivo de variáveis de controle deve ser evitado. Do protocolo HDLC, projetista de protocolos de alto desempenho aprenderam que uma só variável para ACK e Windows não é interessante.

Gerência de Conexão

Controle da Informação

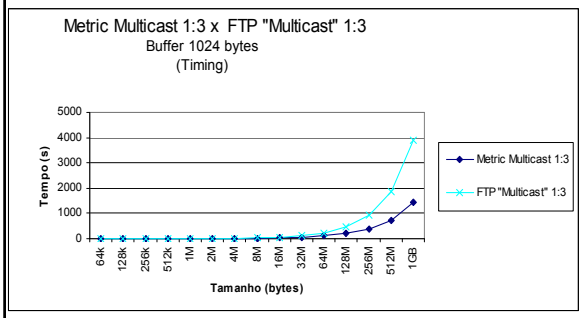
Exemplos são :

• O uso imediato de ACK muitas vezes leva a ***síndrome da janela boba;***

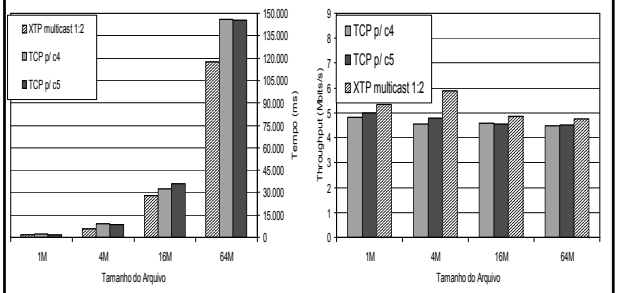
Resultados Experimentais

Vamos a seguir ilustrar parte da pesquisa que estamos desenvolvendo, demonstrando através de comparações um protocolo de alto-desempenho (XTP) e o TCP.

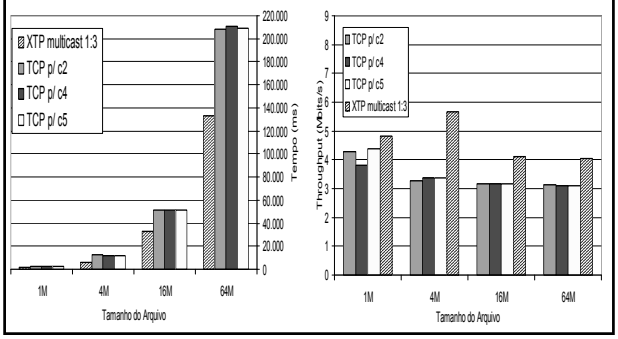
XTP *versus* FTP



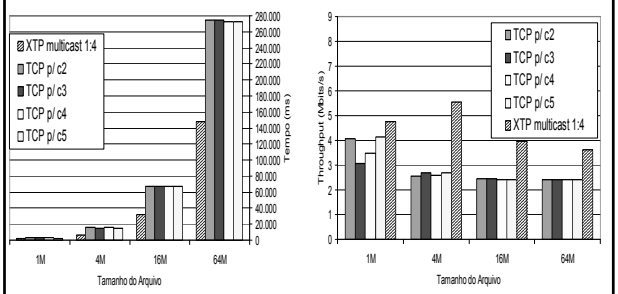
XTP *versus* TCP
Comunicação de 1:2



XTP *versus* TCP
Comunicação de 1:4



XTP *versus* TCP
Comunicação de 1:4



Perguntas ?