



Aula 21: Redes de Interconexão Topologia



Medições de Arquitetura de Interconexão

<i>Interconnect</i>	<i>MPP</i>	<i>LAN</i>	<i>WAN</i>
Example	CM-5	Ethernet	ATM
Topology	“Fat” tree	Line, Bus	Variable, constructed from multistage switches
Connection based?	No	No	Yes
Data Transfer Size	Variable: 4 to 20B	Variable: 0 to 1500B	Fixed: 48B
Store & Forward?	No	n.a.	Yes
Congestion control	At source: Flow control via back pressure	At source: Listen for E-net idle	Rate based via choke packets

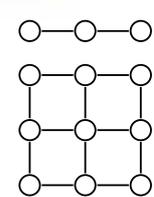


Topologia

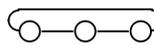
- Estrutura de interconexão
- Determinada por:
 - **Grau**: número de links de um nó
 - **Diâmetro**: número máx de links cruzados entre dois nós
 - **Distância média**: número de links cruzados na média
 - **Bissecção**: número mínimo de links que separam a rede em duas metades
- **Cautela**: desenhos tridimensionais devem ser mapeados a chips e placas bidimensionais
 - Esquemas de interconexão “elegantes” desenhados em papel podem ser difíceis de se implementar



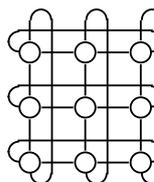
Topologias Importantes



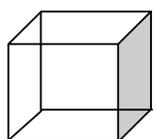
	Type	Degree	Diameter	Ave Dist	Bisection	Diam	Ave D
	1D mesh	≤ 2	$N-1$	$N/3$	1		
	2D mesh	≤ 4	$2(N^{1/2} - 1)$	$2N^{1/2} / 3$	$N^{1/2}$	63	21
	3D mesh	≤ 6	$3(N^{1/3} - 1)$	$3N^{1/3} / 3$	$N^{2/3}$	~30	~10
	nD mesh ($N = k^n$)	$\leq 2n$	$n(N^{1/n} - 1)$	$nN^{1/n} / 3$	$N^{(n-1)/n}$		



	Ring	2	$N / 2$	$N/4$	2		
--	------	---	---------	-------	---	--	--



	2D torus	4	$N^{1/2}$	$N^{1/2} / 2$	$2N^{1/2}$	32	16
	k-ary n-cube ($N = k^n$)	2n	$n(N^{1/n})$ $nk/2$	$nN^{1/n}/2$ $nk/4$	$2k^{n-1}$	15	8 (3D)



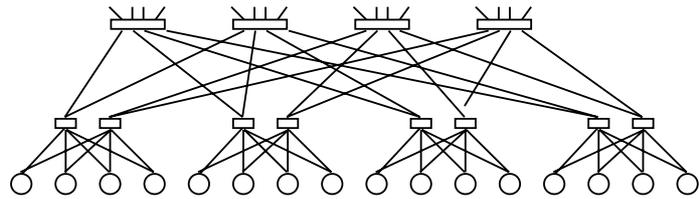
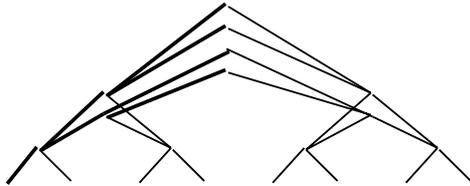
	Hypercube	n	$n = \text{Log}N$	$n/2$	$N/2$	10	5
--	-----------	---	-------------------	-------	-------	----	---



Topologias (continuação)

N = 1024

Type	Degree	Diameter	Ave Dist	Bisection	Diam	Ave D
2D fat tree	4	$\log_2 N$		N		
2D butterfly	4	$\log_2 N$		N/2	20	20

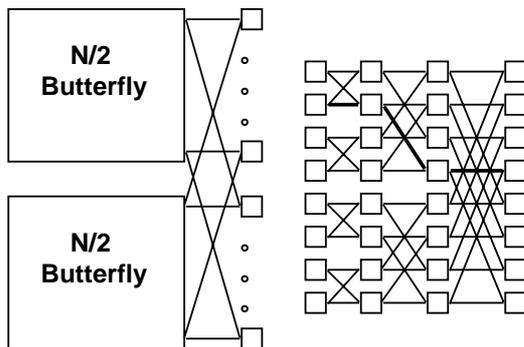


CM-5 Thinned Fat Tree



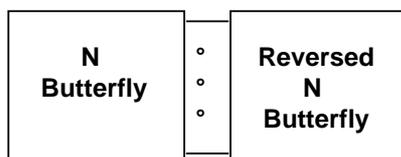
Butterfly

Multi-estágio: nós nas terminações, switches no meio



- Todos os caminhos de mesmo tamanho
- Possui caminho único entre dois nós
- Pode ter conflitos

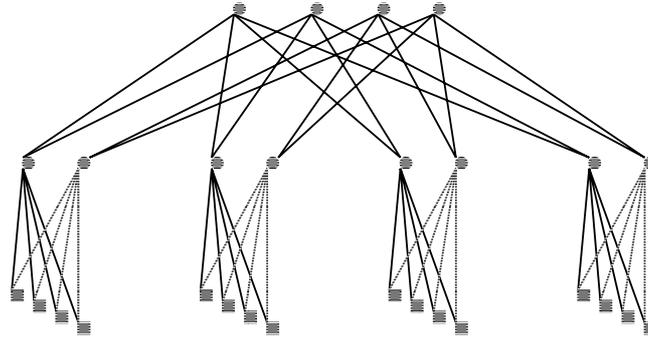
Benes Network



- Roteia todas as combinações sem conflitos



Fat Tree Multi-estágio



- Atribui pacotes aleatoriamente para os diferentes caminhos a fim de se reduzir a carga

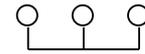


Exemplo

- Em uma rede de 64 nós, assuma que leve 1 unidade de tempo para comunicação entre switches, mas nenhum tempo do switch para o processador. Todos os switches são iguais. Quanto tempo levará para comunicarmos de cada nó para todos os outros nós nas seguintes topologias?
 - Barramento
 - Completamente conectada
 - Anel
 - 2D torus



Exemplo Barramento



- Número de mensagens = $64 \times 63 = 4032$
mensagens
 - Como transferências são sequencias, necessitaremos de 4032 unidades de tempo



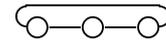
Exemplo Completamente Conectada



- Todas as transferências podem ser feitas em paralelo, logo a comunicação levará uma unidade de tempo



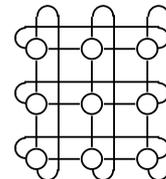
Exemplo Anel



- Cada nó envia mensagem para o nó adjacente com índice mais alto (1-> 2, 2->3, ..., 63->0)
 - Isto leva 1 unidade de tempo
- Cada nó envia mensagem para o nó 2-adjacente com índice mais alto (1->3, 2-> 4, 3->5, ...)
 - Como a mensagem para por dois links, ela levará 2 unidades de tempo
- Só precisamos enviar mensagem até metade dos nós e aí fazemos a mesma operação para os nós de índice mais baixo
 - tempo = $1 + 2 + \dots + 31 + 32 + 31 + \dots + 2 + 1 = 1024$



Exemplo 2D Torus



- Existem 8 linhas e 8 colunas
 - Tempo (mesma linha) = $1 + 2 + 3 + 4 + 3 + 2 + 1 = 16$
 - Tempo (linha debaixo) = $8 \times 1 + \text{Tempo (mesma linha)}$
 - Tempo (2D Torus) = Tempo (mesma linha) + $(8 \times 1 + \text{Tempo (mesma linha)}) + (8 \times 2 + \text{Tempo (mesma linha)}) + \dots + (8 \times 1 + \text{Tempo (mesma linha)})$
 - Tempo (2D Torus) = $8 \times (1 + 2 + 3 + 4 + 3 + 2 + 1 + \text{Tempo (mesma linha)}) = 256$
 - Roteamento pode ser melhorado se não assumirmos duas fases (vertical e horizontal)



Exemplos de Redes

Name	Number	Topology	Bits	Clock	Link	Bisect.	Year
nCube/ten	1-1024	10-cube	1	10 MHz	1.2	640	1987
iPSC/2	16-128	7-cube	1	16 MHz	2	345	1988
MP-1216	32-512	2D grid	1	25 MHz	3	1,300	1989
Delta	540	2D grid	16	40 MHz	40	640	1991
CM-5	32-2048	fat tree	4	40 MHz	20	10,240	1991
CS-2	32-1024	fat tree	8	70 MHz	50	50,000	1992
Paragon	4-1024	2D grid	16	100 MHz	200	6,400	1992
T3D	16-1024	3D Torus	16	150 MHz	300	10,200	1993

Conclusão: Não existe padrão!



Exemplo

- Compare uma rede única de 100Mbps com um switch de 10Mbps. Assuma que um pacote possui na média 250 bytes, a taxa de chegada é de 25000 pacotes por segundo, e que os tempos de chegada de pacotes são exponencialmente distribuídos.



Rede de 100 Mbps

- Taxa de serviço: $100 \times 10^6 / 250 \times 8 = 50000$
- $T_s = 1 / 50000 = 20 \text{ us}$ (tempo servidor)
- $u = r \times T_s = 25000 \times 20 \text{ u} = 0,5$ (utilização)
- $T_q = T_s \times u / (1-u) = 20 \times 0,5 / (1 - 0,5) = 20 \text{ us}$
- Tempo de resposta médio = $20 + 20 = 40 \text{ us}$



Switch c/ 10 redes de 10 Mbps

Seção 6.7

- Taxa de serviço: $10 \times 10^6 / 250 \times 8 = 5000$
- $T_s = 1 / 5000 = 200 \text{ us}$ (tempo servidor)
- $u = r \times T_s / m$
 $= 25000 \times 200 \text{ u} / 10 = 0,5$ (utilização)
- $T_q = T_s \times u / m \times (1-u)$
 $= 200 \times 0,5 / 10 \times (1 - 0,5) = 20 \text{ us}$
- Tempo de resposta médio = $20 + 200 = 220 \text{ us}$



Interconexão Baseada em Conexão vs. Sem Conexão

- Telefone: central faz conexão entre dois aparelhos
 - Uma vez conexão é estabelecida, comunicação pode durar muito tempo
- Podemos utilizar switches para compartilhar linhas de transmissão por longas distâncias longas durante comunicação de diversos aparelhos
 - “**Time division multiplexing**” divide BW da linha de transmissão em um número fixo de slots, cada slot responsável por uma transmissão
- Problema: linhas ficam ocupadas pelo número de linhas conectadas, e não pela quantidade de informação enviada
- Vantagem: BW reservada

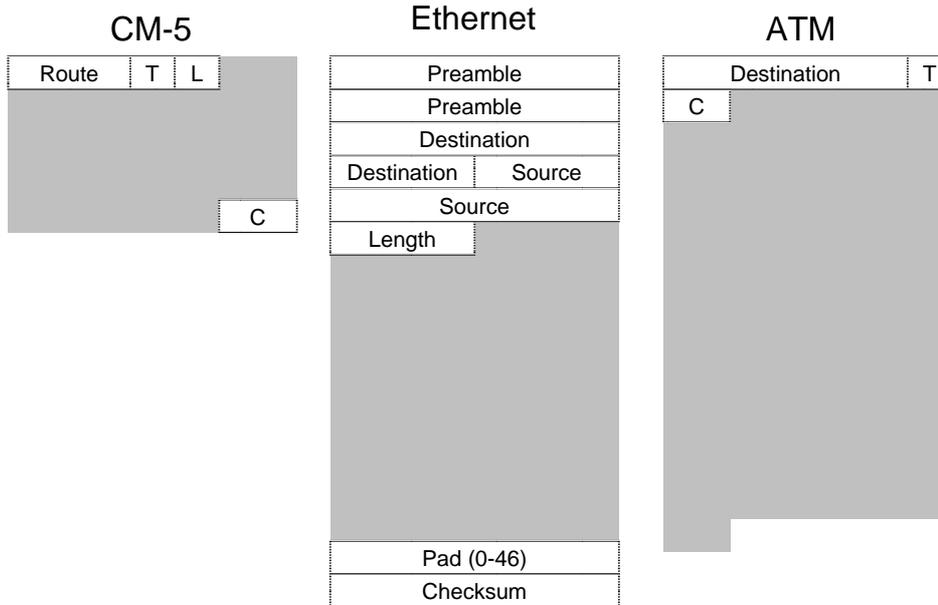


Interconexão Baseada em Conexão vs. Sem Conexão

- **Sem conexão**: cada pacote precisa ter endereço de destino
 - Cada pacote pode ser roteado para o destino baseado em seu endereço (parecido com sistema postal)
 - Barramentos do tipo *split phase* utiliza envio de pacotes
 - Multiplexação baseada na carga



Formato de Pacotes



- Campos: Destino, Checksum(C), Comprimento(L), Tipo(T)
- Tamanho de Dado/Cabeçalho em bytes: (4 a 20)/4, (0 a 1500)/26, 48/5



Controle de Congestionamento

- Redes com chaveamento de pacotes não reservam BW; por isso ter congestionamento
- Solução: recusar pacotes a entrar na rede até congestionamento ser resolvido
- Opções:
 - **Packet discarding**: Se um pacote chega no switch e não existe espaço no buffer, pacote é descartado
 - **Flow control**: entre pares de tx/rx; rx avisa tx quando pode receber mais pacotes
 - **Back-pressure**: utiliza canais separados para avisar quando parar
 - **Window**: permite tx enviar N pacotes de cada vez
 - **Choke packets**: aka “rate-based”; Cada pacote recebido por switch ocupado é retornado para fonte marcado como pacote recusado. Fonte reduz tráfego para destino por uma % fixa (ATM Forum)



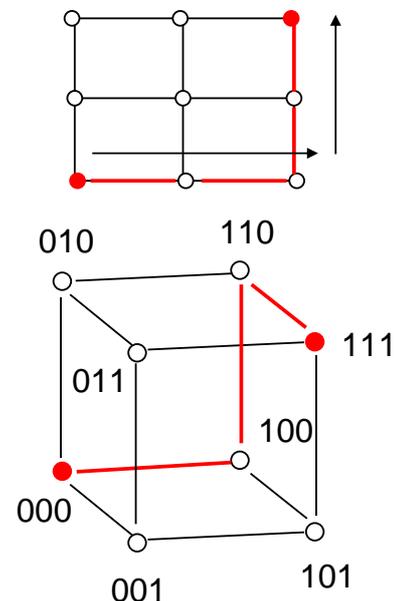
Store and Forward vs. Cut-Through

- **Store-and-forward**: cada switch aguarda pacote ser recebido para reenviá-lo para o próximo switch
- **Cut-through** ou **worm hole**: switch examina o cabeçalho do pacote e decide para aonde enviar o pacote, e então envia-o imediatamente
- **Vantagem**: Latência é dada por
 - Store and Forward:
número de switches intermediários X tamanho do pacote
 - Cut-Through:
tempo para primeira parte do pacote ser utilizada para negociar rota + tamanho do pacote ÷ BW da rede de interconexão



Roteamento

- **Determinístico**—segue rota pré-especificada
 - mesh: roteamento por dimensão
 - $(x_1, y_1) \rightarrow (x_2, y_2)$
 - primeiramente $\Delta x = x_2 - x_1$,
 - depois $\Delta y = y_2 - y_1$,
 - hypercube: roteamento por aresta
 - $X = x_0x_1x_2 \dots x_n \rightarrow Y = y_0y_1y_2 \dots y_n$
 - $R = X \text{ xor } Y$
 - Precisa buscar rota em que $x_i \text{ xor } y_i = 1$ (diferem)
 - tree: ancestral comum
- **Adaptativo**—baseia-se no estado da rede de interconexão (e.g., congestionamento)





Questões Práticas

Tipo	MPP	LAN	WAN
Exemplo	CM-5	Ethernet	ATM
Padrão	Não	Sim	Sim
Tolerância falhas	Não	Sim	Sim
<i>Hot Insert</i>	Não	Sim	Sim

- Padrão: requerido para WAN, LAN!
- Tolerância a falhas: Pode um nó falhar e o sistema continuar a funcionar?
- *Hot Insert*: Se a interconexão pode suportar falhas, ela também pode continuar a operação enquanto um novo nó é adicionado a rede?

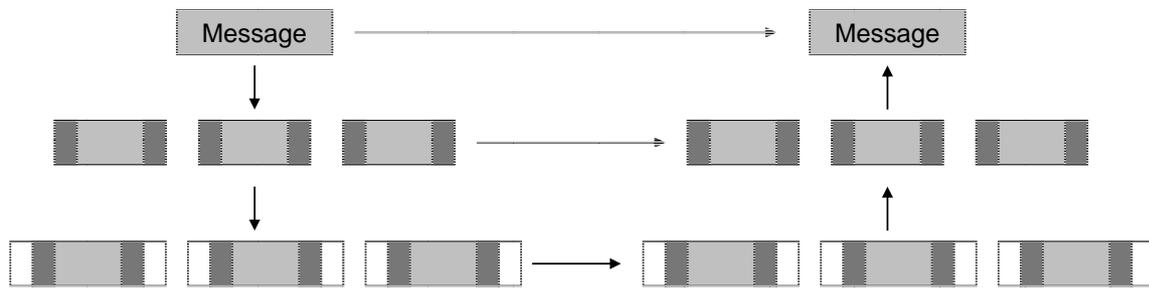


Protocolos: Interface de HW/SW

- *Internetworking*: permite computadores em redes independentes e incompatíveis a se comunicarem de modo confiável e eficiente;
 - *Enabling technologies*: Padrões de SW que permitem comunicação confiável em redes não confiáveis
- *Transmission Control Protocol/Internet Protocol (TCP/IP)*
 - Base da internet



Protocolo



- Chave para família de protocolos é que comunicação ocorre logicamente no mesmo nível do protocolo (*peer-to-peer*), mas é implementada como serviços no nível mais baixo
- Cada nível adicional aumenta a latência
 - TCP/IP sobre Ethernet e redes de alta velocidade



Resumo: Interconexões

- Comunicação entre computadores
- Implementação: comprimento, meio de tx
- Performance: overhead, latência, BW
- Topologias: muitas para escolher, mas overhead de SW fazem todas elas parecerem iguais, mas com diferentes custos