

Barramentos de I/O RAID

Tendências de Interconexão

- Interconexão = interfaceia computador com os componentes do sistema de computação
- Interfaces em hardware de alta velocidade + protocolos
- *Networks, channel, backplanes*

	Network	Channel	Backplane
Distance	>1000 m	10 - 100 m	1 m
Bandwidth	10 - 100 Mb/s	40 - 1000 Mb/s	320 - 1000+ Mb/s
Latency	high (>ms)	medium	low (<μs)
Reliability	low Extensive CRC	medium Byte Parity	high Byte Parity
Baseados em mensagens via de dados estreita arbitragem distribuída		←→	Mapeados em memória via de dados larga arbitragem centralizada

Arquiteturas de Backplane

Metric	VME	FutureBus	MultiBus II	SCSI-I
Bus Width (signals)	128	96	96	25
Address/Data Multiplexed?	No	Yes	Yes	na
Data Width	16 - 32	32	32	8
Xfer Size	Single/Multiple	Single/Multiple	Single/Multiple	Single/Multiple
# of Bus Masters	Multiple	Multiple	Multiple	Multiple
Split Transactions	No	Optional	Optional	Optional
Clocking	Async	Async	Sync	Either
Bandwidth, Single Word (0 ns mem)	25	37	20	5, 1.5
Bandwidth, Single Word (150 ns mem)	12.9	15.5	10	5, 1.5
Bandwidth Multiple Word (0 ns mem)	27.9	95.2	40	5, 1.5
Bandwidth Multiple Word (150 ns mem)	13.6	20.8	13.3	5, 1.5
Max # of devices	21	20	21	7
Max Bus Length	.5 m	.5 m	.5 m	25 m
Standard	IEEE 1014	IEEE 896	ANSI/IEEE 1296	ANSI X3.131

Observações:

SCSI channel funciona como um barramento

FutureBus funcional como um channel (*disconnect/reconnect*)

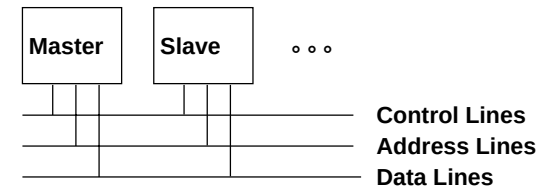
Interconexões Baseadas em Barramentos

- Barramento: meio de comunicação compartilhado entre subsistemas
 - Baixo custo: conjunto de fios compartilhados
 - Versatilidade: Fácil de adicionar novos dispositivos e periféricos
- Desvantagem
 - Limitação de velocidade de comunicação causa gargalo para throughput
- Velocidade de barramento é limitado por fatores físicos
 - Comprimento
 - Carga (número de dispositivos conectados)

Interconexões Baseadas em Barramentos

- Dois tipos genéricos:
 - Barramentos de I/O: longo, dispositivos de diversas velocidades e tipos e seguem um padrão de barramento
 - Barramentos CPU/memória: alta velocidade, maximiza BW entre CPU e memória
 - Para baixar custos, sistemas mais antigos combinavam os dois barramentos
- Transação de barramento é dividida em duas fases
 - Envio de endereço
 - Envio/recepção de dados

Protocolos de Barramento



Multibus: 20 endereço, 16 dados, 5 controle, 50ns Pausa

Bus Master: controla o barramento e inicia a transação

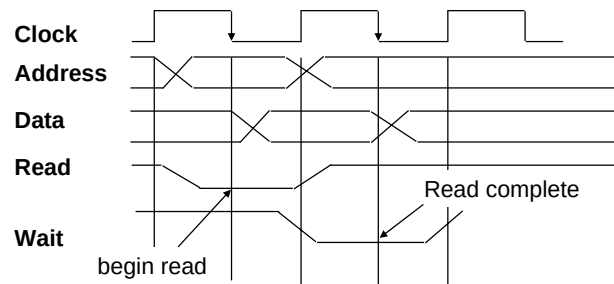
Bus Slave: ativado pela transação

Protocolo de comunicação: especificação de sequência de eventos e temporizações para transferir informação

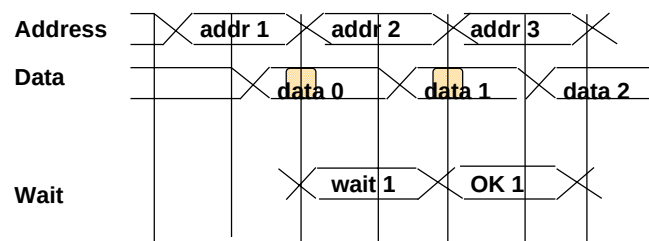
Transferências assíncronas: linhas de controle (req., ack.) servem para controlar a transação

Transferências síncronas: transação é controlada por um clock

Protocolos Síncronos

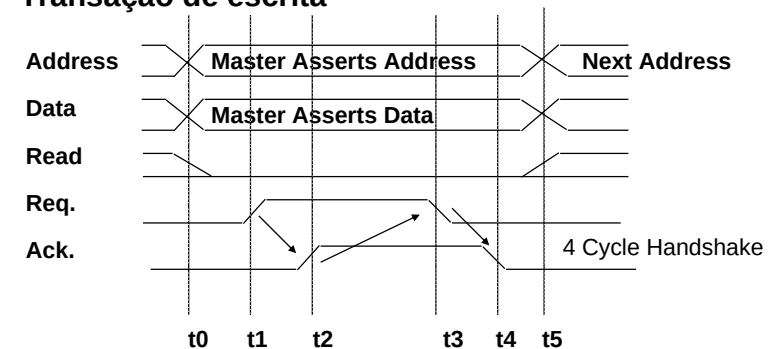


Pipelined/Split transaction Bus Protocol



Handshake Assíncrono

Transação de escrita



t0 : Mestre obtém controle e envia endereço, direção e dado
Aguarda um tempo para *target* decodificar endereço

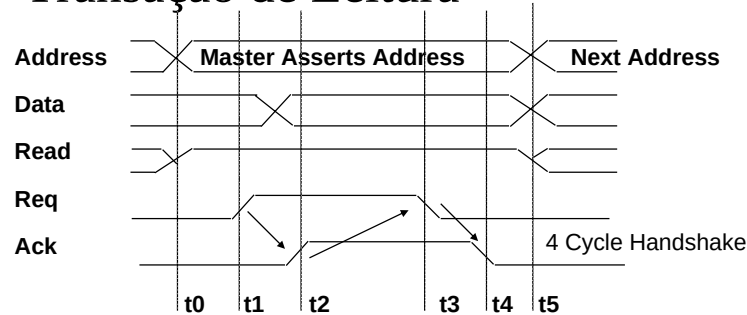
t1: Mestre asserts req.

t2: Escravo asserts ack, indicando recebimento de dado

t3: Mestre libera req

t4: Escravo libera ack

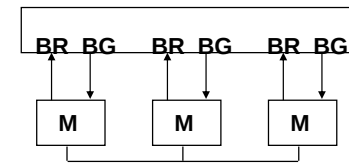
Transação de Leitura



- t0: Mestre obtém controle e envia endereço, direção e dado
Aguarda um tempo para *target* decodificar endereço
- t1: Mestre *asserts* req.
- t2: Escravo *asserts* ack, indicando transmissão do dado
- t3: Mestre libera req, dado recebido
- t4: Escravo libera ack

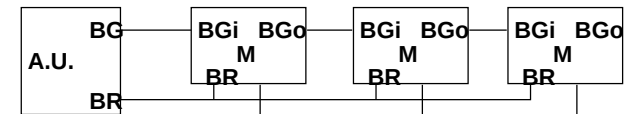
Arbitragem de Barramento

Parallel (Centralized) Arbitration

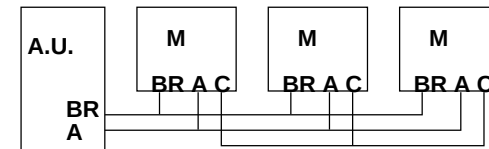


BR = *Bus Request*
BG = *Bus Grant*

Serial Arbitration (daisy chaining)



Polling



Opções de Barramento

Opção	Alta performance	Baixo custo
Bus width	Endereço e dados em linhas separadas	Multiplexação das linhas de endereços e dados
Data width	Largo é mais rápido (32 bits)	Estreito é mais barato (8 bits)
Transfer size	Múltiplas palavras diminuem overhead	Uma palavra é mais simples
Bus masters	Múltiplos (requer arbitragem)	Único (sem arbitragem)
Split transaction?	Sim—envio e recepção aumenta BW (requer múltiplos mestres)	Não—conexão contínua é mais barato e possui menor latência
Clocking	Síncrono	Assíncrono

Barramentos 1990 (P&H, 1st Ed)

	VME	FutureBus	Multibus II	IPI	SCSI
Signals	128	96	96	16	8
Addr/Data mux	no	yes	yes	n/a	n/a
Data width	16 - 32	32	32	16	8
Masters	multi	multi	multi	single	multi
Clocking	Async	Async	Sync	Async	either
MB/s (0ns, word)	25	37	20	25	1.5 (asyn) 5 (sync)
150ns word	12.9	15.5	10	=	=
0ns block	27.9	95.2	40	=	=
150ns block	13.6	20.8	13.3	=	=
Max devices	21	20	21	8	7
Max meters	0.5	0.5	0.5	50	25
Standard	IEEE 1014	IEEE 896.1	ANSI/IEEE ANSI X3.129	ANSI X3.131	

VME

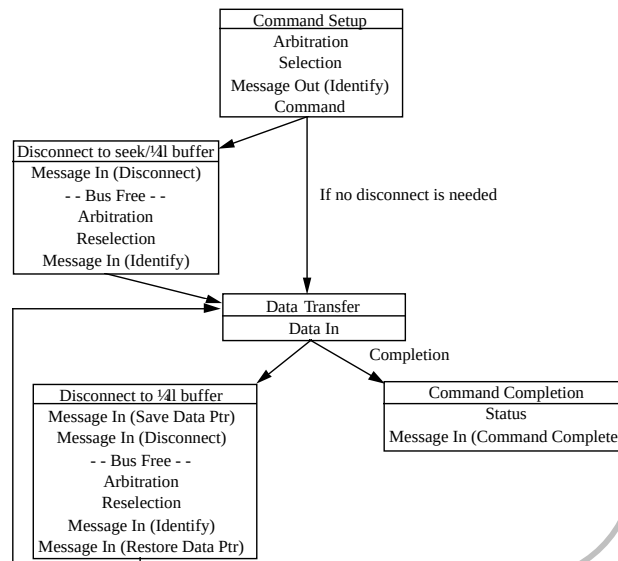
- 3 conectores de 96 pinos
- 128 pinos definidos no padrão, restante definido pelo usuário
 - 32 pinos para endereço
 - 32 pinos para dados
 - 64 pinos para comando e linhas de alimentação

SCSI: Small Computer System Interface

- Até 8 dispositivos comunicam-se em um barramento a velocidades de até 4-5 MBytes/sec
- SCSI-2 aumenta velocidade para até 20 MB/sec
- Dispositivos podem ser escravos (“target”) ou mestres (“initiator”)
- Protocolo: seqüência de fases, durante as quais ações são tomadas pelo controlador e dispositivos SCSI
 - **Bus Free**: Nenhum dispositivo está acessando o barramento
 - **Arbitration**: barramento está livre, logo múltiplos dispositivos podem requisitar o barramento utilizando prioridade fixa por endereço
 - **Selection**: Informa qual target será utilizado (**Reselection** se desconectado)
 - **Command**: initiator lê bytes de comando da memória do host e envia comandos para target
 - **Data Transfer**: dados enviados entre initiator e target
 - **Message Phase**: mensagem enviada entre initiator e target (*identify, save/restore data pointer, disconnect, command complete*)
 - **Status Phase**: target, antes de completar comando

SCSI : Arquitetura

peer-to-peer protocols
initiator/target
linear byte streams
disconnect/reconnect



Ex. de Barramentos (P&H, 2nd Ed)

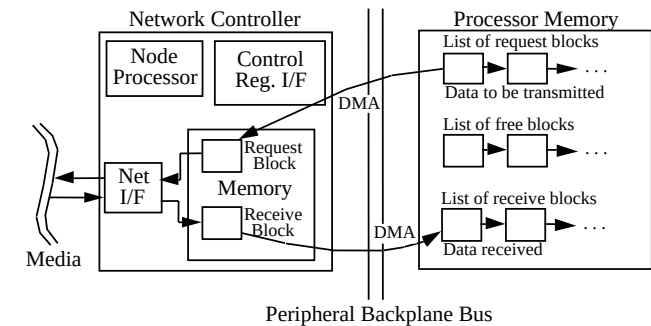
Bus	SBus	TurboChannel	MicroChannel	PCI
Originator	Sun	DEC	IBM	Intel
Clock Rate (MHz)	16-25	12.5-25	async	33
Addressing	Virtual	Physical	Physical	Physical
Data Sizes (bits)	8,16,32	8,16,24,32	8,16,24,32,64	8,16,24,32,64
Master	Multi	Single	Multi	Multi
Arbitration	Central	Central	Central	Central
32 bit read (MB/s)	33	25	20	33
Peak (MB/s)	89	84	75	111 (222)
Max Power (W)	16	26	13	25

1993 Barramentos de Memória p/ MP

Bus	Summit	Challenge	XDBus
Originator	HP	SGI	Sun
Clock Rate (MHz)	60	48	66
Split transaction?	Yes	Yes	Yes?
Address lines	48	40	??
Data lines	128	256	144 (parity)
Data Sizes (bits)	512	1024	512
Clocks/transfer	4	5	4?
Peak (MB/s)	960	1200	1056
Master	Multi	Multi	Multi
Arbitration	Central	Central	Central
Addressing	Physical	Physical	Physical
Slots	16	9	10
Busses/system	1	1	2
Length	13 inches	12? inches	17 inches

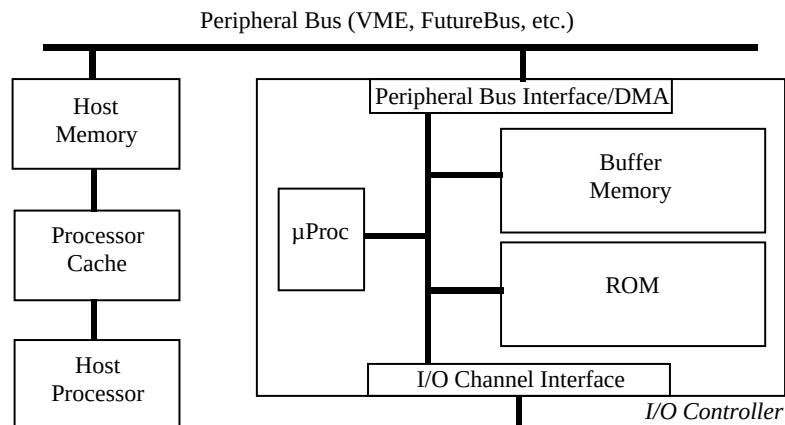
Redes de Comunicação

Limitação de performance é devida a transferências de memória, overhead de OS, e não protocolos



- Filas de envio e recepção na memória do computador
- Controlador de rede copia dados via DMA
- Host não participa do controle
- Causa interrupção ao final da transação

Arquitetura de Controladores de I/O

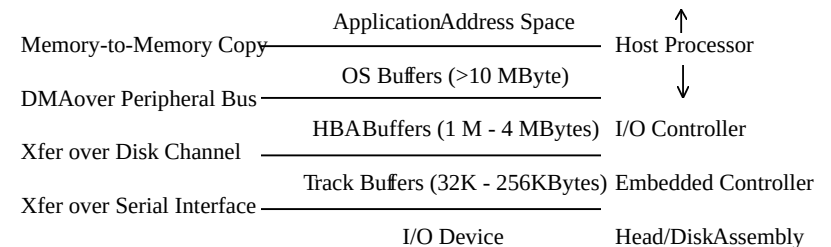


Request/response block interface

Backdoor access to host memory

Fluxo de Dados de I/O

Limite de performance: cópias múltiplas, hierarquia complexa



Armazenamento sobre Rede

Decreasing Disk Diameters

14" » 10" » 8" » 5.25" » 3.5" » 2.5" » 1.8" » 1.3" » ...

Rede fornece interfaces lógicas e físicas bem definidas: por que não separar CPU de sistema de armazenamento?

Serviço de Armazenamento sobre rede de alta velocidade

Network File Services

OS suportando acesso remoto

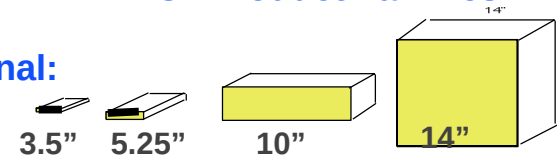
3 Mb/s » 10Mb/s » 50 Mb/s » 100 Mb/s » 1 Gb/s » 10 Gb/s
rede capaz de sustentar transferências de alto BW

Increasing Network Bandwidth

Vantagens de Fabricação de Disk Arrays

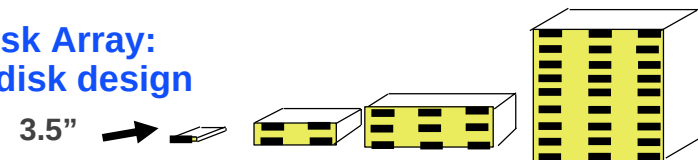
Disk Product Families

Conventional:
4 disk designs



Low End → High End

Disk Array:
1 disk design



Troca de Pequeno # de Discos Grandes por Grande # de Discos pequenos!

	IBM 3390 (K)	IBM 3.5" 0061	x70
Data Capacity	20 GBytes	320 MBytes	23 GBytes
Volume	97 cu. ft.	0.1 cu. ft.	11 cu. ft.
Power	3 KW	11 W	1 KW
Data Rate	15 MB/s	1.5 MB/s	120 MB/s
I/O Rate	600 I/Os/s	55 I/Os/s	3900 I/Os/s
MTTF	250 KHrs	50 KHrs	??? Hrs
Cost	\$250K	\$2K	\$150K

Disk Array tem o potencial para

- taxa alta de dados e I/O
- alta densidade, baixa potência
- reliability baixa

Redundant Arrays of Disks

• Redundância leva a alta disponibilidade

Eventualmente discos falharão

Conteúdo pode ser reconstruído por dados armazenados c/ redundância no array

- Capacidade reduz devido a redundância
- BW reduz devido aos múltiplos acessos que serão necessários

Técnicas:

- Espeelhamento (alto custo em capacidade)
- Hamming Codes (custo muito alto)
- Paridade & Códigos Reed-Solomon

Reliability do Array

- Reliability de N discos = Reliability de 1 disco $\div$ N

50,000 Horas $\div$ 70 discos = 700 horas

MTTF do sistema de discos: cai de 6 anos para 1 mês!

- Arrays sem redundância não tem reliability para serem úteis!

Redundant Arrays of Disks (RAID)

- Espelhamento de disco**

Cada disco é duplicado em disco de proteção

escrita lógica = duas escritas físicas

100% de overhead

- Parity Data Bandwidth Array**

Paridade calculada horizontalmente

um disco com alto BW

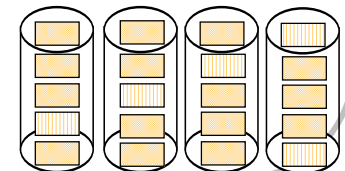
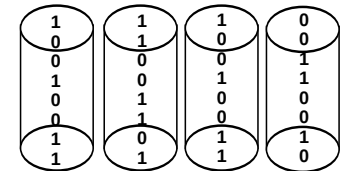
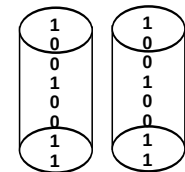
- High I/O Rate Parity Array**

Blocos de paridade intercalados

leituras e escritas independentes

escrita lógica = 2 leituras + 2 escritas

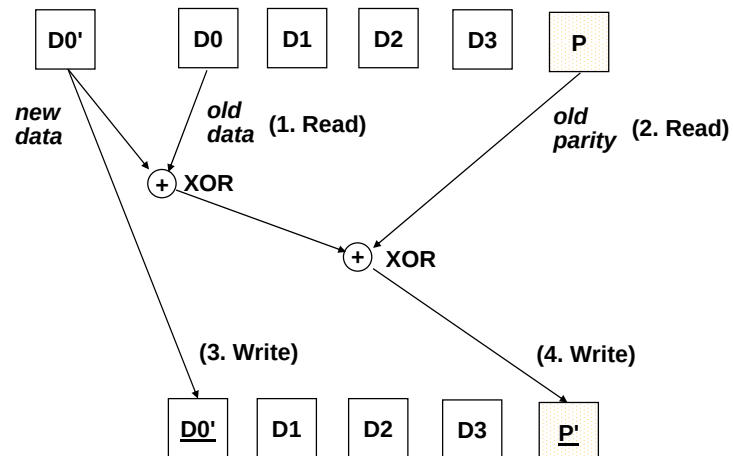
paridade



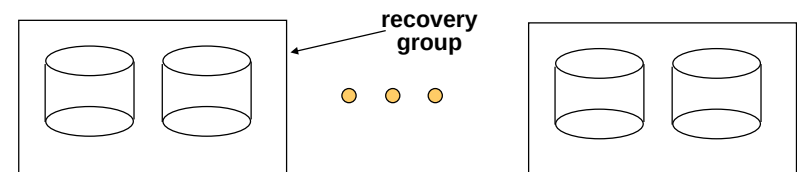
Problemas com Arrays de Discos:

RAID-5: Small Write Algorithm

1 escrita lógica = 2 leituras + 2 escritas



RAID 1: Espelhamento



- Cada disco é copiado inteiramente em imagem
- Disponibilidade é alta

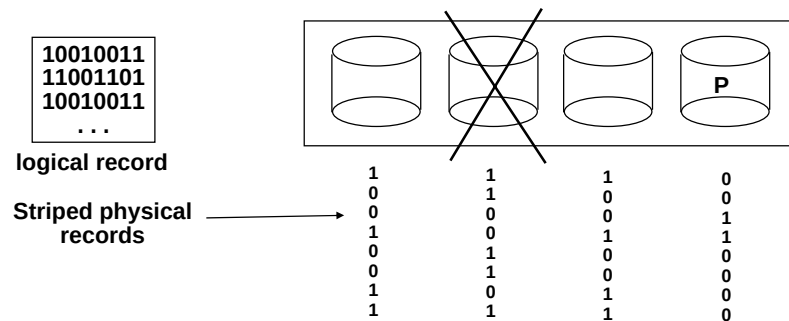
- Sacrifício em BW: escrita lógica = duas escritas físicas

- Leituras podem ser otimizadas

- cara: 100% de overhead de capacidade

Destinados para I/O rate alto com alta disponibilidade

RAID 3: Disco de Paridade



- Paridade calculada para todo o grupo para evitar falhas
33% a mais sobre custo da capacidade
- Se todos os braços sincronizados logicamente => alta capacidade, alta velocidade

Aplicações: Científica, PDI

RAID 5+: High I/O Rate Parity

