Armazenamento e Sistema E/S

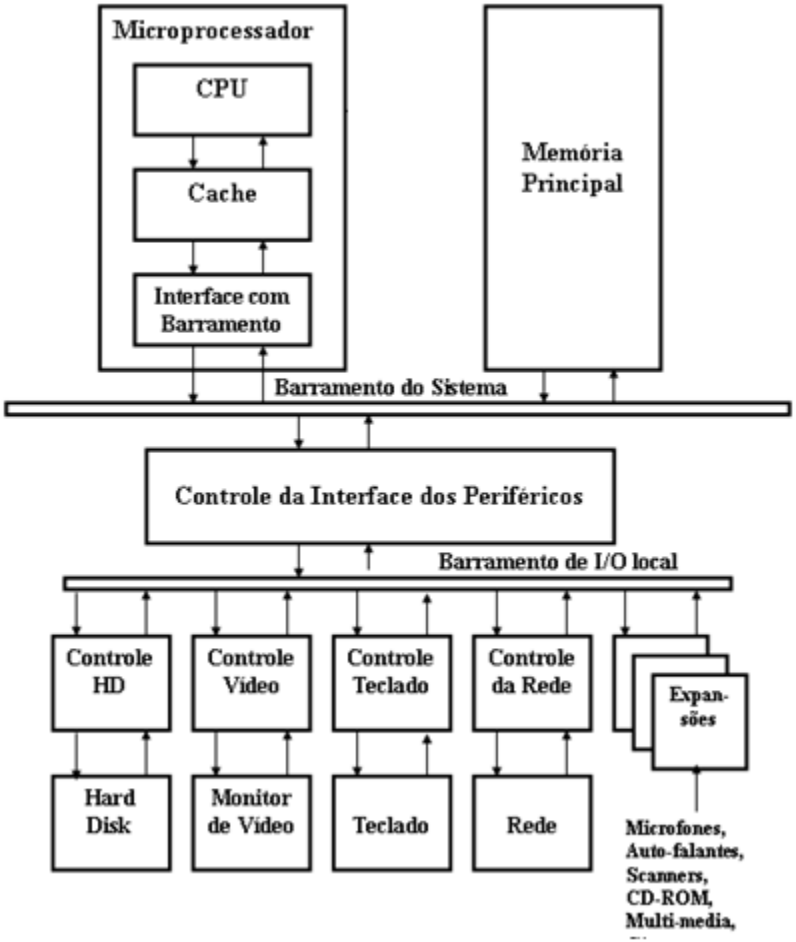
CES-25 – Arquiteturas para Alto Desmpenho
Prof. Paulo André Castro
pauloac@ita.br
Sala 110 – Prédio da Computação
www.comp.ita.br/~pauloac
IEC - ITA

Conteúdo

- Armazenamento: Discos
- Servidores de E/S
  - Desempenho
  - Confiabilidade e Disponibilidade
- Sistemas RAID
- Armazenamento: Memória Flash
- Barramentos

Armazenamento

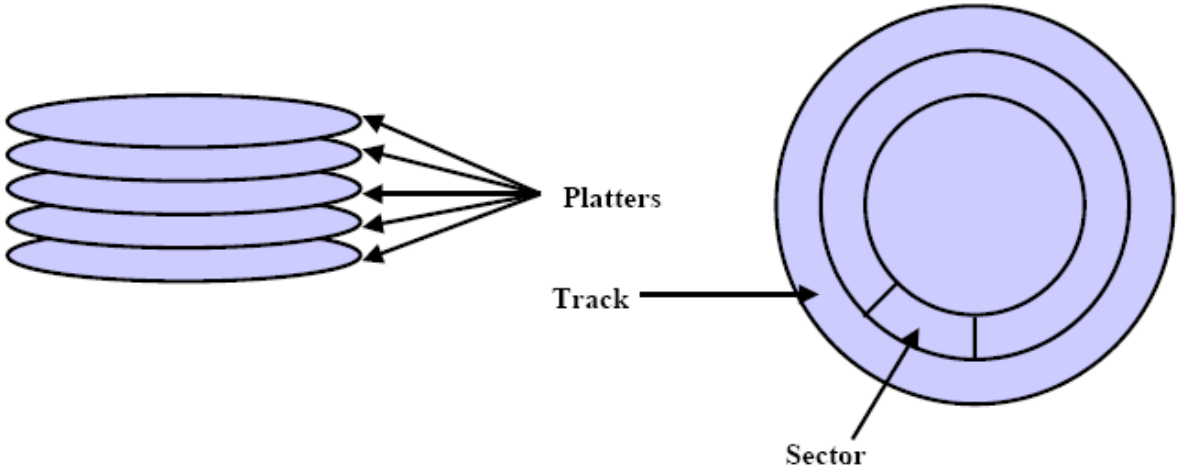
- A memória não volátil pode ser vista como parte do sistema de hierarquia de memória
- ..ou como parte do sistema de E/S pois invariavelmente é conectada aos barramentos de E/S e não ao barramento da memória principal
- Como Armazenar?
  - Discos Magnéticos
  - Memória Flash

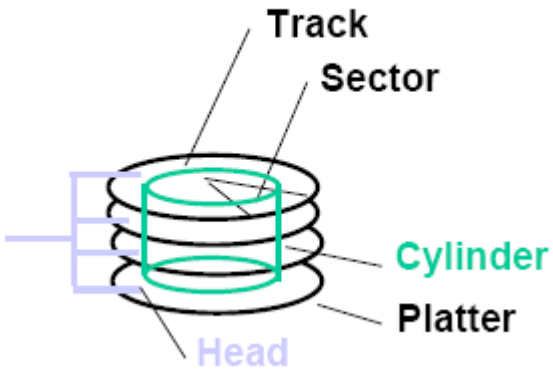


Discos Magnéticos

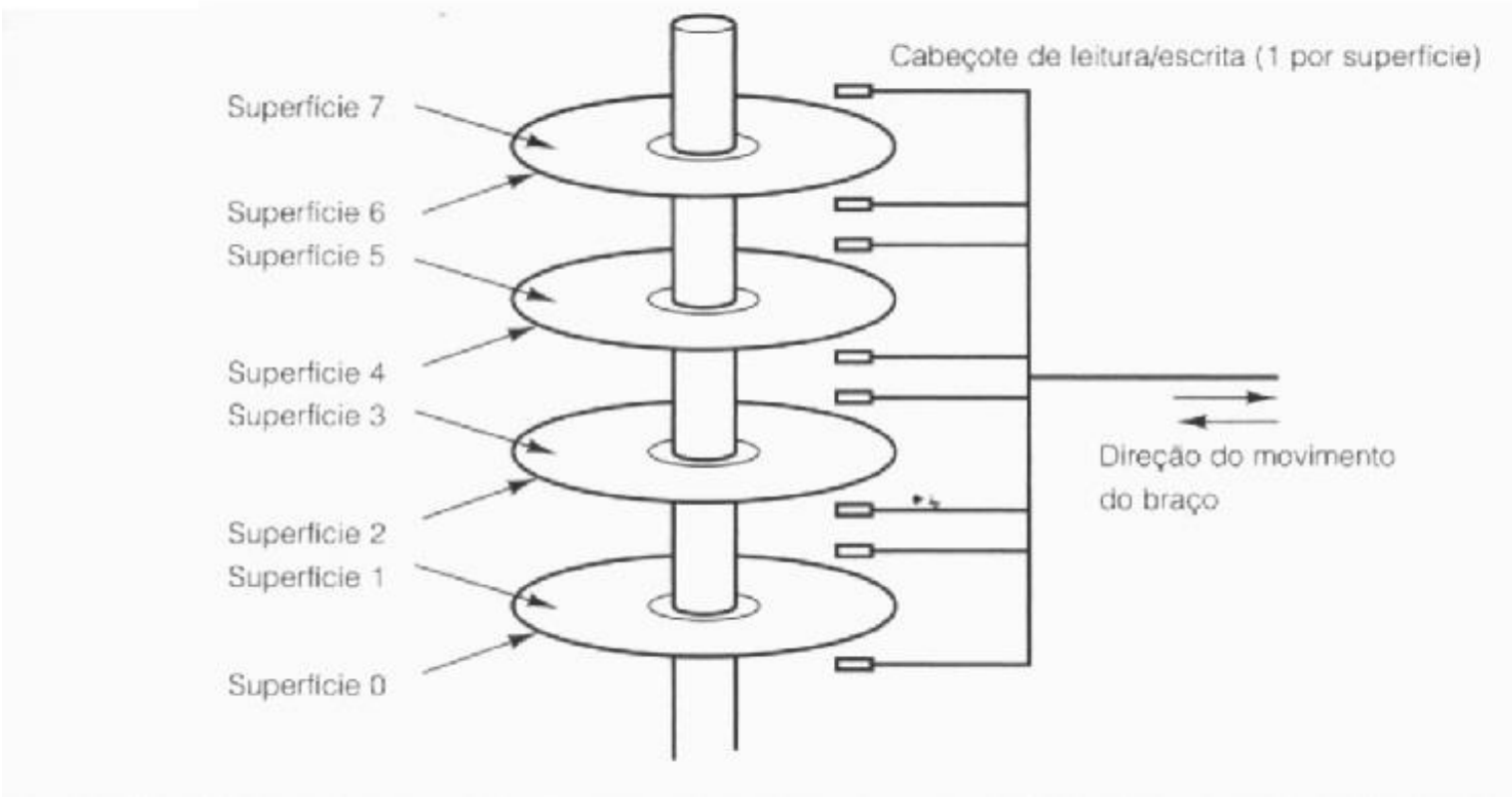
- **Propósito:**
  - Armazenamento não-volátil
  - Grande, barato e lento
  - Nível mais baixo na hierarquia de memórias
- Usados no passado também como dispositivo para transporte físico de dados *(floppy disks)*
- Baseia-se em um disco rotativo coberto com uma superfície magnética
- Usam uma cabeça(head) de leitura/escrita para acessar as informações
- Vantagens dos Discos rígidos (HD) sobre Floppy disks:
  - Como os disco são rígidos(metal ou vidro) podem ser maiores
  - Maior densidade porque podem ser controlados com mais precisão
  - Maior taxa de transferência porque podem rodar mais rápido
  - Podem ter mais de um “disco” (platter)

Organização de um Disco





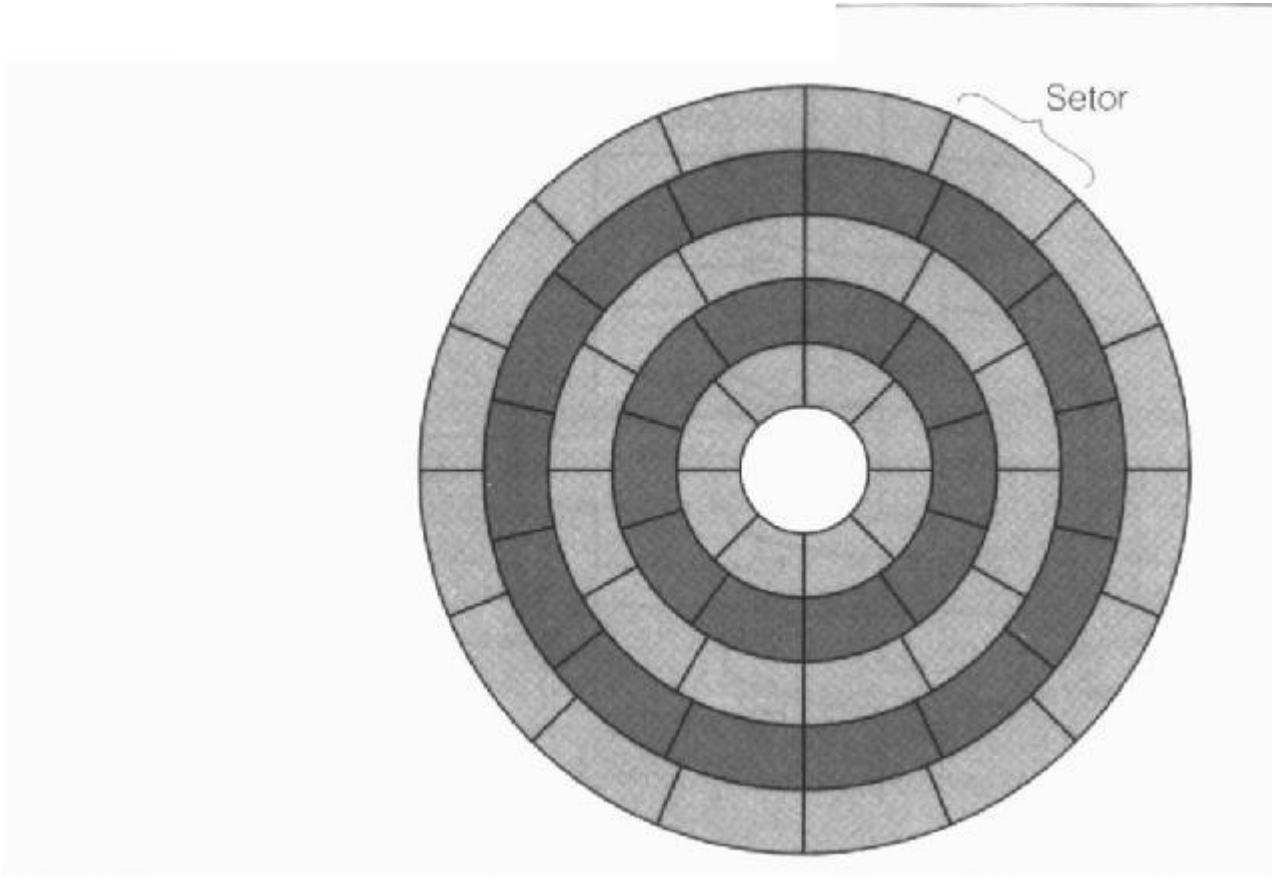
Discos e superfícies



Trilha e Setores

- Números Típicos (dependem do tamanho do disco)
  - 5.000 a 30.000 trilhas(tracks) por superfície
  - 100 a 500 setores(sectors) por trilha
    - Setor: menor unidade que pode ser lida
- Geralmente, todas as trilhas tinham o mesmo número de setores
  - Logo: setores tem tamanhos físicos distintos
- Atualmente, discos tem trilhas com diferentes números de setores para garantir discos com maior capacidade

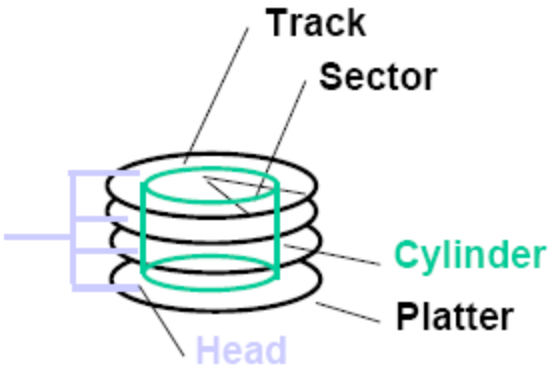
Trilhas e setores



- Há menos setores nas trilhas internas

Disco Magnético

- Cilindros: Todas as trilhas sobre a cabeça de leitura/ escrita das superfícies.
- Processo de Leitura/Escrita
  1. Posicionar o braço na trilha correta (seek time)
  2. Roda o disco até que o setor esteja sobre a cabeça de leitura (rotational latency)
  3. Ler ou gravar (transferir) um bloco de dados (transfer time)



Desempenho de Discos Magnéticos

- Seek Time: na faixa de 5 a 12 ms
  - Soma de todos os tempos de buscas/Número de Buscas
  - Devido à “localidade” o seek time real pode ser apenas 25% a 30% do tempo divulgado pelos fabricantes.

Latência Rotacional

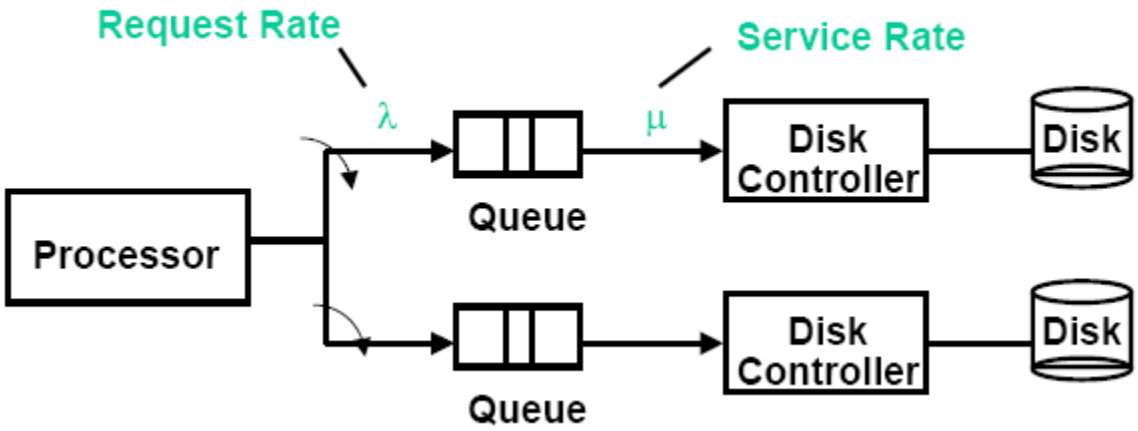
- Rotational Latency:
  - Período de rotação do disco: 3,600 a 10,000 RPMs (16ms a 0,4ms por rotação)
  - Latência média: Tempo para percorrer metade do disco (8ms a 0,2 ms)

- Latência Rotacional=0,5 * Periodo de rotação = 0,5/X RPM = 0,5/ (X*60*RPS)

Desempenho de Discos Magnéticos

- Tempo de Transferência: fatores relevantes
  - Tamanho da transferência(1 setor): 1KB/setor
  - Taxa de Transferência: 3 a 65MB/s
    - Velocidade de Rotação: 3600 a 15000 RPM
    - Densidade de bits: bits/polegada
    - Diâmetro do disco: 1,0 a 3,5 polegadas
- Valores típicos de Transfer Time: 0,01 a 0,03ms/setor

Tempo de Acesso ao Disco



Disk Access Time = Queuing Delay + Controller Time + Seek time + Rotational Latency + Transfer time

Exercício: Calcule o tempo de acesso ao Disco.

512 byte sector, rotate at 5400 RPM, advertised seeks is 12 ms, transfer rate is 4 MB/sec, controller overhead is 1 ms, queue idle, so no service time

Solução

Disk Access Time = Seek time + Rotational Latency + Transfer time + Controller Time + Queuing Delay

= 12 ms + 0.5 / 5400 RPM + 0.5 KB / 4 MB/s + 1 ms + 0 ms

= 12 ms + 0.5 / 90 RPS + 0.125 / 1024 s + 1 ms + 0 ms

= 12 ms + 5.5 ms + 0.1 ms + 1 ms + 0 ms

= 18.6 ms

- **If real seeks are 1/3 advertised seeks, then its 10.6 ms, with rotation delay contributing to 50% of the time!**

Evolução dos Discos Magnéticos

- **Evolução**: aumento do número de bits por polegada quadrada.
- **Custos**: Queda acentuada de US$ 100.000/GB em 1984 para menos de 0.5$/GB em 2012
- **Desempenho**:
  - Aumento de RPM de 3.600 RPM na década de 80 para proximo a 10.000 RPM nos anos 2000, não continuou a crescer devido a problemas com alta velocidade de rotação....
  - Juntamente com o aumento de densidade tem-se obtido por volta de 40% de ganho de desempenho por ano.

Year	Improve/yr	Doubling time
< 1988	29%	3 yrs
< 1996	60%	1.5 yrs
< 2001	100%	1 yr

Discos não devem falhar

- Discos diferem dos demais níveis de hierarquia de memória, porque são não-voláteis
- E são também o nível mais baixo. Não há um onde buscar no computador se o dado não estiver no disco.
- Portanto, discos não devem falhar...mas todo hardware falha.

Discos Redundantes

RAID : Redundant Array of Inexpensive Disks

- Múltiplos acessos são feitos simultaneamente
- Dados são “espalhados” pelos vários discos
  - Stripping
  - Mirroring

- Stripping
  - Dados seqüenciais são alocados logicamente em discos separados para aumentar desempenho
- Mirroring
  - Dados são copiados em discos idênticos (espelhos) para aumentar disponibilidade
- Caracterísiticas
  - Latência não necessariamente é reduzida
  - Disponibilidade é maior através da adição de discos redundantes
    - Informação perdida é reconstruída através da informação redundante

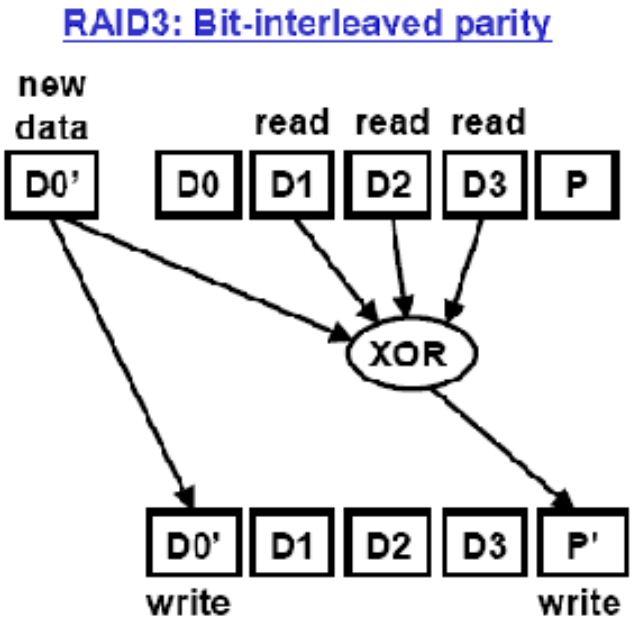
RAID
Confiabilidade X Disponibilidade

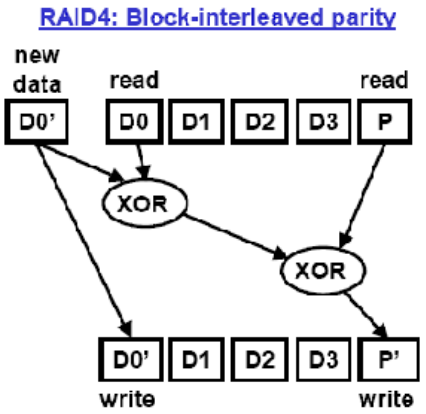
- Confiabilidade é menor
  - Mais discos, maior probabilidade de falha
- Entretanto, disponibilidade é maior
  - Falhas não levam necessariamente a indisponibilidade

Níveis de RAID

- RAID 0:
  - Não redundante, porém mais eficiente. Não se recupera de falhas
- RAID 1:
  - Redundante e capaz de se recuperar de uma falha. Entretanto, usa o dobro de discos do RAID 0
- RAID 2 : Não tem implementações comerciais
- RAID 3,4 e 5: 1 disco de check para vários discos de dados, capacidade de recuperação para uma falha. Todos baseados em operações XOR.
- RAID 6: Dois discos de check e capacidade de sobreviver a duas falhas.

RAID 3 e RAID 4



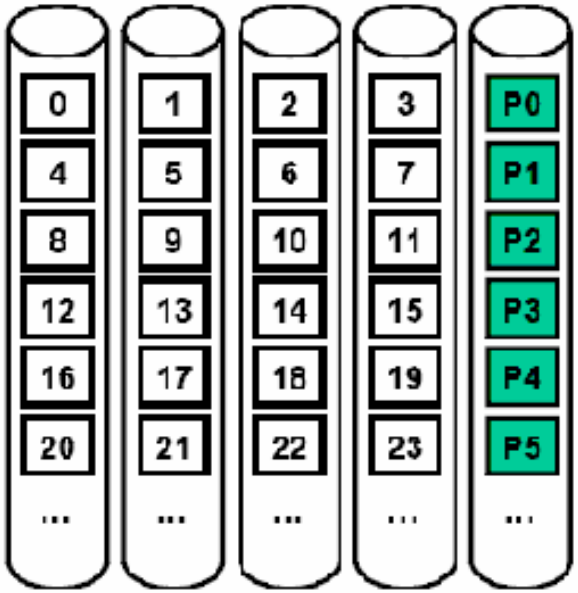


In RAID3, small updates involve reading the unmodified disks to update the parity disk as well as the disk to be modified

RAID4 minimizes the interaction with unmodified disks so as to allow independent accesses to occur

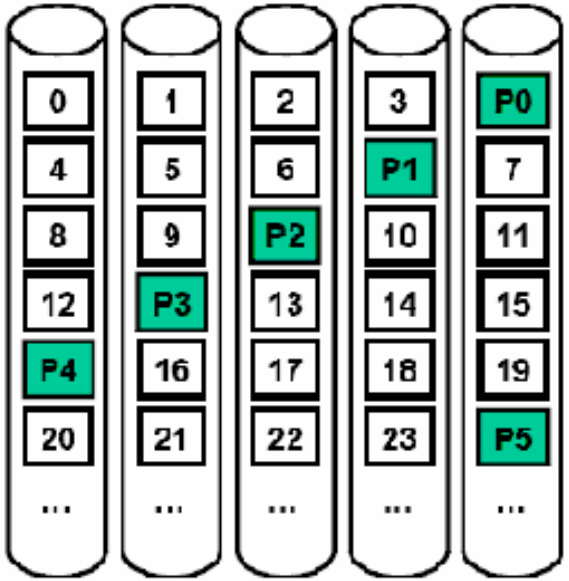
RAID 4 e RAID 5

RAID4: Block-Interleaved parity



In RAID4, sequential writes need to be serialized due to centralization of parity blocks

RAID5: Distributed block-interleaved parity



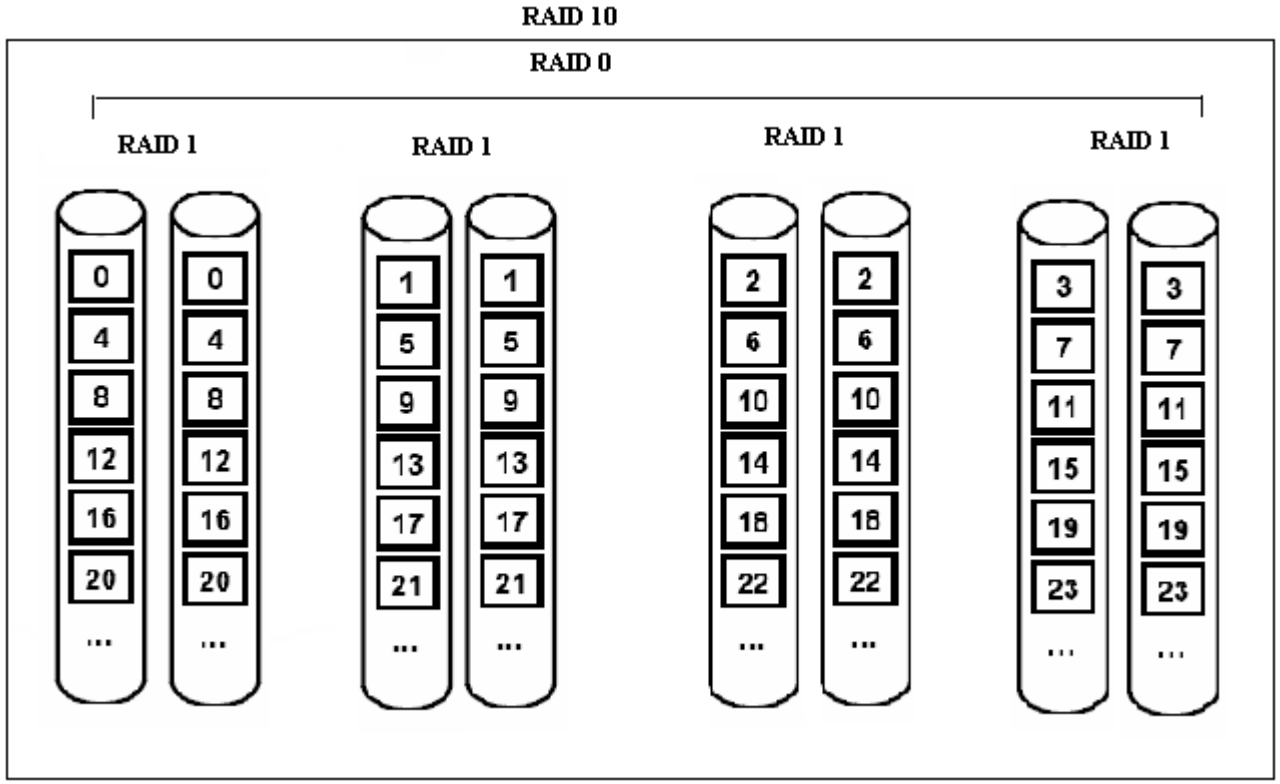
RAID5 allows simultaneous accesses to proceed e.g. writes to blocks 5 and 8

Sistemas RAID

- RAID 6: Guarda duas paridades (P+Q), com isso pode recuperar-se de até duas falhas. Utiliza dois discos para paridade.
- Buscam maior disponibilidade do sistema de disco
  - Menor confiabilidade: maior probabilidade de falha
  - Falhas de disco não necessariamente levam a falhas do sistema de disco
  - Sistema redundante em disco e com capacidade de recuperação mesmo sem reinicialização do computador (HotSwap)

Outras Variações RAID

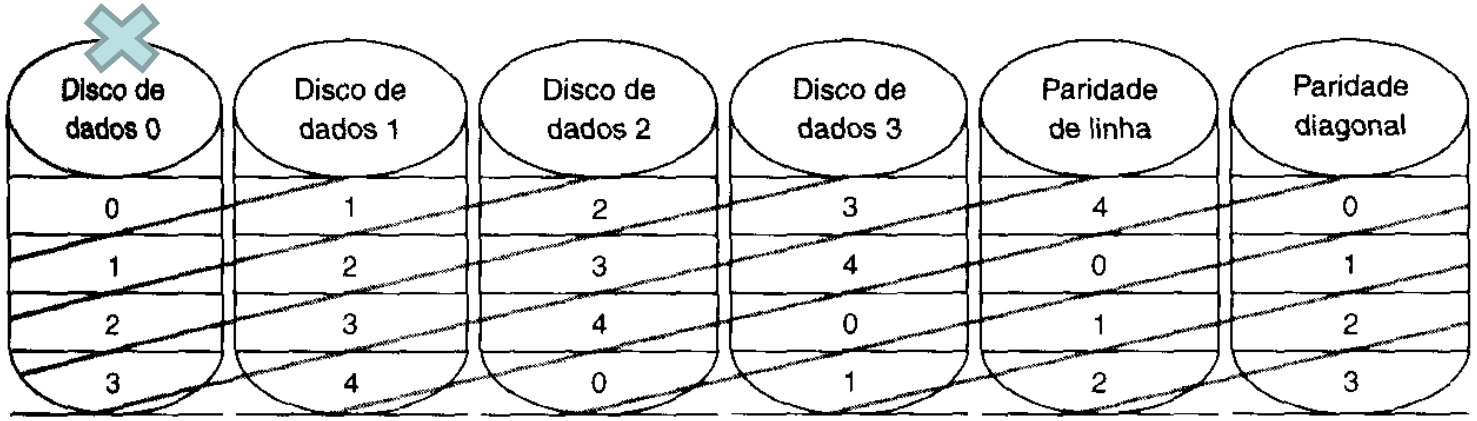
- RAID 10: RAID 1 (Mirroring) + 0 (Stripping)
  - Exemplo: 4 pares de disco, cada par espelhado e os pares dividindo dados
- RAID 01 ?



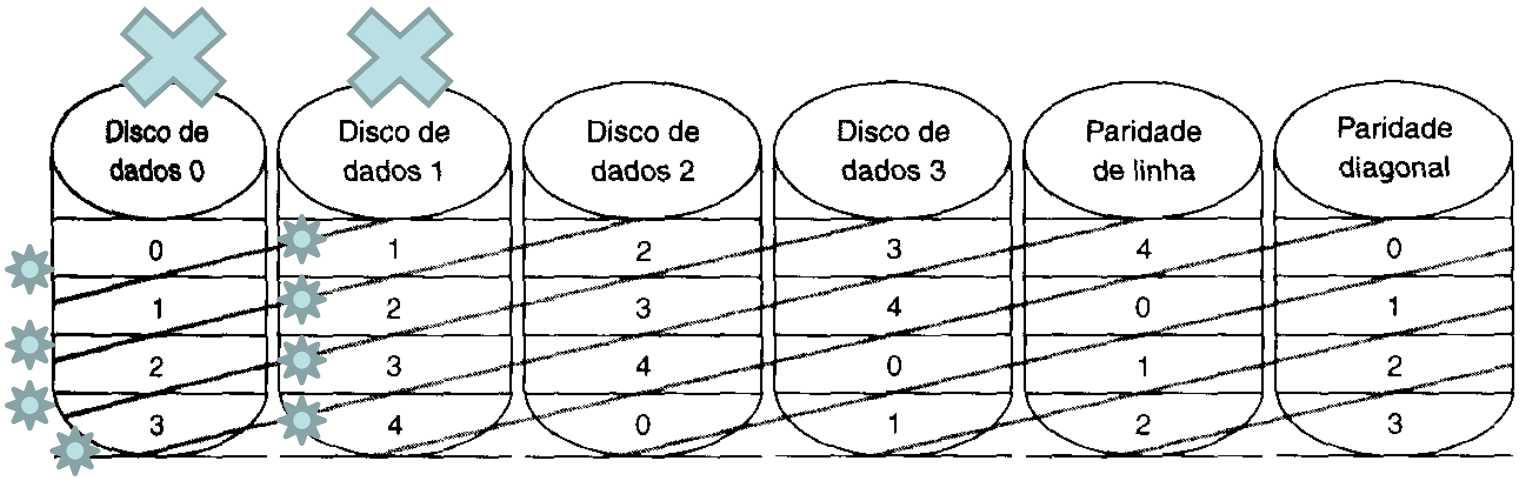
Detalhamento RAID 6

- Dois discos de paridades:
  - Um disco de paridade construído por linha como no RAID 4
  - Um disco de paridade construído por diagonal
- Cada diagonal exclui um disco, portanto mesmo se falharem dois discos será possível recuperar um bloco, recuperado um bloco pode-se recuperar o segundo através da linha

Detalhamento do RAID 6



Recuperação de Falha Dupla no RAID 6



Memória Flash

- Tecnologia similar a tradicionais EEPROM, maior capacidade de memória por chip
- Baixo consumo de energia
- Tempo de acesso de leitura mais lento que DRAM porém muito mais rápido que discos
  - Em 2010, uma transferência de 256 bytes de Flash levaria em torno de 6.5Microsegundos e 1000 vezes mais em disco
  - Para escritas, a DRAM pode ser de 10 a 100 vezes mais rápida...
- Gravação exige que seja deleção prévia dos dados
  - Primeiro apaga-se um bloco de memória e depois grava-se

Flash NOR e NAND

- As primeiras memórias flash, (Flash NOR) era um concorrente direto das tradicionais EEPROM sendo aleatoriamente endereçável.
- Depois de algum tempo, surgiram as memórias flash NAND que oferecem maior densidade de armazenamento , mas só pode ser lida em blocos, pois elimina a fiação necessária para o acesso aleatório
  - Flash NAND é muito mais barata por gigabyte e muito mais comum que flash NOR
  - Memórias Flash NOR são tipicamente usadas em sistemas básicos de entrada e saída (BIOS)

Memória Flash

- Em 2010, preço/GB era 2$/GB para flash, 40$/GB para SDRAM e 0.09$/GB para discos
- Em 2016, preço/GB era 0,3$/GB para flash, 7$/GB para SDRAM e 0.06$/GB para discos
- porém há desgaste da memória flash nas escritas, normalmente limitado a algo entre 100K e 1M gravações...
- O tempo de vida é expandido através da distribuição uniforme das escritas através dos blocos
- Eliminou o discos flexíveis e está eliminando os discos rígidos em sistemas móveis....Solid State Disks

Memória Flash

- Qual o Tempo de leitura e gravação de dados de 64KB em memória Flash e disco magnético ? (dados de (Hennensy,Patterson, 2014))
  - Memória Flash:
    - 25MicroS/leitura de 1 bloco(2K)
    - 250 MicroS para gravação e 1,5ms para apagar bloco (2048B)
  - Disco: Overhead de controlador: 1ms
    - 3600RPM
    - 12ms de seek time anunciado(real igual a 1/3)

Memória Flash

- Flash
  - Leitura: 64KB/2KB* 25 MicroS= **0,8ms**
  - Gravação: 64KB/2KB *250MicroS + 64KB/2KB*1500MicroS = **56 ms**
- Disco
  - Leitura/Gravação
    - 12ms/3 + 0,5/3600RPM + 64KB/4,2MB/s +0,1ms =
    - **27,3ms**

Servidores de E/S (Clusters)

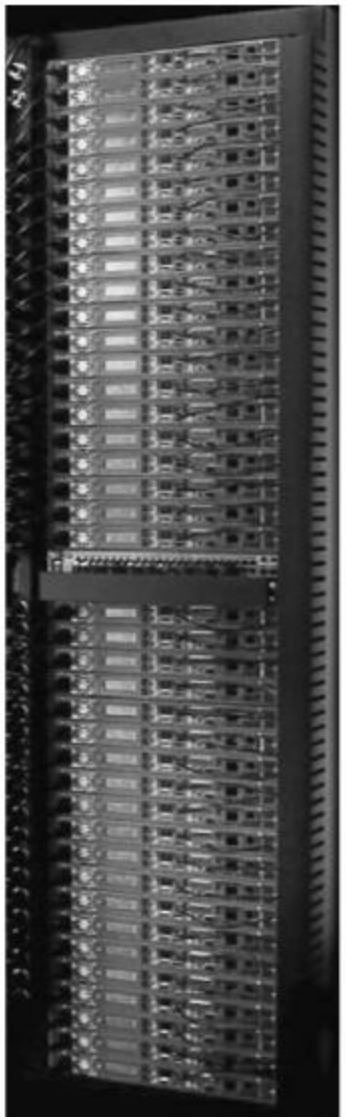
- Avaliando **custo**, **desempenho** e **confiabilidade** de um sistema projetado para fornecer **alto desempenho de I/O**

- Exemplo: Rack VME T-80 utilizado no sistema Internet Archive (projeto iniciado em 1996 que visa fazer o registro histórico da Internet ao longo do tempo…)

Building Blocks of Clusters…

- Typical building blocks of Cluster: 1U **server** (left), 7´ **rack** with Ethernet switch (right). Barroso and Urs Holzle (2009),
- 1U= 1,75 inches = 4,45cm, 7’ = 7 feet = 84 inches = 2,13 m (provides 48 U)

Rack VME T-80 from Capricornian Systems



Basic building block is a 1U storage node called the PetaBox GB2000 from Capricorn Technologies.

It uses four 500 GB Parallel ATA (PATA) disk drives, 512 MB of DDR266 DRAM, one 10/100/1000 Ethernet interface, and a 1 GHz C3 processor from VIA (80x86 instruction set)

This node dissipates about 80 watts in typical configurations.

40 nodes of the GB2000s fit in a standard VME rack, which gives the rack 80 TB of raw capacity.

The 40 nodes are connected together with a 48-port 10/100/1000 switch, and it dissipates about 3 KW. The limit is usually 10 KW per rack in computer facilities,

Desempenho

- Estimar o desempenho em IOPS (operações de E/S por segundo) de um rack T-80. Dados adicionais e hipóteses simplificadoras:
  - The VIA processor, 512 MB of DDR266 DRAM, ATA disk controller, power supply, fans, and enclosure cost $500.
  - Each of the four 7200 RPM Parallel ATA drives holds 500 GB, has an average time seek of 8.5 ms, transfers at 50 MB/sec from the disk, and costs $375. The PATA link speed is 133 MB/sec
  - The performance of the VIA processor is 1000 MIPS.

Hipóteses

- The ATA controller adds 0.1 ms of overhead to perform a disk I/O.
- The operating system uses 50,000 CPU instructions for a disk I/O.
- The network protocol stacks use 100,000 CPU instructions to transmit a data block between the cluster and the external world.
- The average I/O size is 16 KB for accesses to the historical record via the Wayback interface, and 50 KB when collecting a new snapshot.

Problema: Custo de IOPS

- **Evaluate the cost per I/O per second (IOPS) of the 80 TB rack**. Assume that every disk I/O requires an average seek and average rotational delay. Assume that the workload is evenly divided among all disks and that all devices can be used at 100% of capacity; that is, the system is limited only by the weakest link, and it can operate that link at 100% utilization. Calculate for both average I/O sizes.

- Solução: Calcular a capacidade de cada “link” do sistema e utilizar o menor como limitante

Blade (PetaBox GB2000)

- Número máximo de IOPS por processador, memória

$$\text{Maximum IOPS for CPU} = \frac{1000 \text{ MIPS}}{50{,}000 \text{ instructions per I/O} + 100{,}000 \text{ instructions per message}} = 6667 \text{ IOPS}$$

The maximum performance of the memory system is determined by the memory bandwidth and the size of the I/O transfers:

$$\text{Maximum IOPS for main memory} = \frac{266 \times 8}{16 \text{ KB per I/O}} \approx 133{,}000 \text{ IOPS}$$

$$\text{Maximum IOPS for main memory} = \frac{266 \times 8}{50 \text{ KB per I/O}} \approx 42{,}500 \text{ IOPS}$$

Barramento de E/S

$$\text{Maximum IOPS for the I/O bus} = \frac{133 \text{ MB/sec}}{16 \text{ KB per I/O}} \approx 8300 \text{ IOPS}$$

$$\text{Maximum IOPS for the I/O bus} = \frac{133 \text{ MB/sec}}{50 \text{ KB per I/O}} \approx 2700 \text{ IOPS}$$

- Como cada blade (GB2000), tem dois barramentos no máximo teremos 16.600 IOPS para 16KB de E/S e 5400 IOPS para 50KB de E/S

Do Barramento para o controlador

Now it's time to look at the performance of the next link in the I/O chain, the ATA controllers. The time to transfer a block over the PATA channel is

$$\text{Parallel ATA transfer time} = \frac{16\ \text{KB}}{133\ \text{MB/sec}} \approx 0.1\ \text{ms}$$

$$\text{Parallel ATA transfer time} = \frac{50\ \text{KB}}{133\ \text{MB/sec}} \approx 0.4\ \text{ms}$$

Adding the 0.1 ms ATA controller overhead means 0.2 ms to 0.5 ms per I/O, making the maximum rate per controller

$$\text{Maximum IOPS per ATA controller} = \frac{1}{0.2\ \text{ms}} = 5000\ \text{IOPS}$$

$$\text{Maximum IOPS per ATA controller} = \frac{1}{0.5\ \text{ms}} = 2000\ \text{IOPS}$$

Disco

The next link in the chain is the disks themselves. The time for an average disk I/O is

$$\text{I/O time} = 8.5\text{ ms} + \frac{0.5}{7200\text{ RPM}} + \frac{16\text{ KB}}{50\text{ MB/sec}} = 8.5 + 4.2 + 0.3 = 13.0\text{ ms}$$

$$\text{I/O time} = 8.5\text{ ms} + \frac{0.5}{7200\text{ RPM}} + \frac{50\text{ KB}}{50\text{ MB/sec}} = 8.5 + 4.2 + 1.0 = 13.7\text{ ms}$$

Therefore, disk performance is

$$\text{Maximum IOPS (using average seeks) per disk} = \frac{1}{13.0\text{ ms}} \approx 77\text{ IOPS}$$

$$\text{Maximum IOPS (using average seeks) per disk} = \frac{1}{13.7\text{ ms}} \approx 73\text{ IOPS}$$

or 292 to 308 IOPS for the four disks.

Rede

The final link in the chain is the network that connects the computers to the outside world. The link speed determines the limit:

$$\text{Maximum IOPS per 1000 Mbit Ethernet link} = \frac{1000 \text{ Mbit}}{16\text{K} \times 8} = 7812 \text{ IOPS}$$

$$\text{Maximum IOPS per 1000 Mbit Ethernet link} = \frac{1000 \text{ Mbit}}{50\text{K} \times 8} = 2500 \text{ IOPS}$$

- Lembre-se: B= bit*8

Em resumo, os limites são:

- Processador: 6667 IOPS
- Memória: 133.000 IOPS (16KB per I/O) e 42500 (50KB per I/O)
- Barramento: 16.600 IOPS e 5400 IOPS
- Controlador: 5000 a 2000 IOPS
- Discos: 308 e 292 (Quatro discos)
- Rede: 7812 e 2500 IOPS
- Logo, o gargalo de desempenho é determinado pelos discos
  - O rack com 40 GB2000 operará a 40x308=12.320 IOPS ou 40x292=11680 IOPS
  - Se o switch não tiver capacidade de operar com 12320*16K*8= 1.6Gbit/s ou 11680*50K*8=4.7Gbit/s, então ele seria o gargalo.

- Logo, o gargalo de desempenho é determinado pelos discos
  - O rack com 40 GB2000 operará a 40x308=12.320 IOPS ou 40x292=11680 IOPS
  - Se o switch não tiver capacidade de operar com 12320*16K*8= 1.6Gbit/s ou 11680*50K*8=4.7Gbit/s. Então ele seria o gargalo.
  - Assumimos que as 8 portas extras de 1000Mbit conectam o rack para o resto do mundo então, elas devem ser capazes de suportar o IOPS máximo dos 160 discos do rack.

- Custo,
  - 40 x (500+4*375)+3000+1500 (rack) = $84.500
  - Os discos representam quase 60% do total
  - O custo por TB de armazenamento é aprox. $1000 (10 a 15 mais baixo que a versão anterior do Internet Archive)
  - O custo por IOPS é em torno de $7

Confiabilidade/Disponibilidade

- Como definir confiabilidade para o cluster T-80 (ou outro sistema computacional) ?
- Como calcular ?
- e disponibilidade ?

Disponibilidade e Confiabilidade

- Confiabilidade é uma medida de realização de um serviço sem falhas. Uma forma de medir confiabilidade seria o tempo médio para uma falha (MTTF)
- Disponibilidade é uma medida da realização de um serviço sem interrupção do mesmo. Uma forma de medir seria:
  - Disponibilidade = MTTF/(MTTF+MTTR), onde
- MTTR: tempo médio para reparo.
- Outro valor comumente utilizado é tempo médio entre falhas, MTBF. Onde MTBF = MTTF+MTTR
- Taxa de falha = 1/MTTF (falhas por unidade de tempo)

Exemplo

- Considerando um subsistema de disco com os seguintes componentes, calcule o MTTF do sistema:
  - 10 discos, cada um com 1.000.000 horas MTTF
  - 1 controladora SCSI, 500.000h de MTTF
  - 1 fonte de alimentação, 200.000h de MTTF
  - 1 cabo SCSI, 1.000.000 h de MTTF
  - Conside falhas independentes e MTTF constante ao longo do tempo.

Exemplo

- Taxa de falha = 10 * 1/1.000.000 + 1/500.000 + 1/200.000+1/1.000.000 = 23/1.000.000

- $MTTF_{sistema}$ = 1/T.F. = 43.500 horas

Confiabilidade do Cluster

- Como calcular a MTTF do Rack?
- Vamos usar as seguintes informações sobre MTTF dos componentes

  - CPU/memory/enclosure MTTF is 1,000,000 hours.
  - PATA Disk MTTF is 125,000 hours.
  - PATA controller MTTF is 500,000 hours.
  - Ethernet Switch MTTF is 500,000 hours.
  - Power supply MTTF is 200,000 hours.
  - Fan MTTF is 200,000 hours.
  - PATA cable MTTF is 1,000,000 hours.

MTTF e taxa de falha

$$\text{Failure rate} = \frac{40}{1{,}000{,}000} + \frac{160}{125{,}000} + \frac{40}{500{,}000} + \frac{1}{500{,}000} + \frac{40}{200{,}000} + \frac{40}{200{,}000} + \frac{80}{1{,}000{,}000}$$

$$= \frac{40 + 1280 + 80 + 2 + 200 + 200 + 80}{1{,}000{,}000 \text{ hours}} = \frac{1882}{1{,}000{,}000 \text{ hours}}$$

The MTTF for the system is just the inverse of the failure rate:

$$\text{MTTF} = \frac{1}{\text{Failure rate}} = \frac{1{,}000{,}000 \text{ hours}}{1882} = 531 \text{ hours}$$

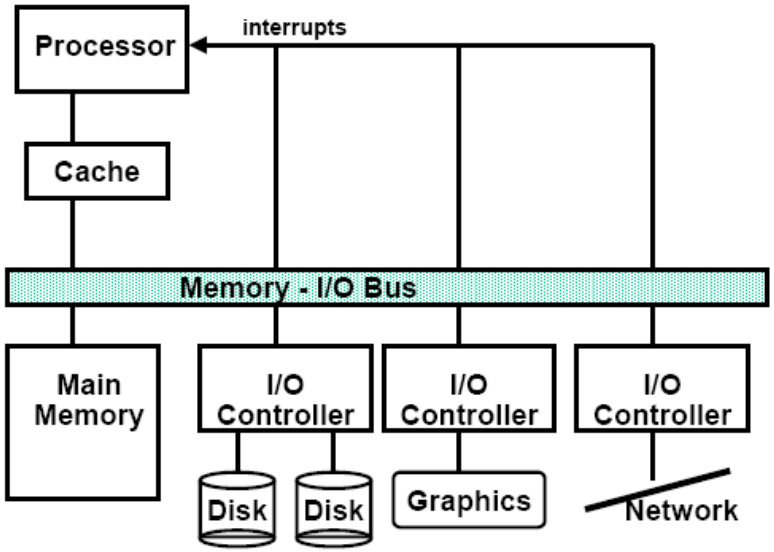
- Obs.: Cabo PATA permite conectar dois HDs ATA
- 3 semanas = 504 horas… 531= 22 dias e 3 horas…

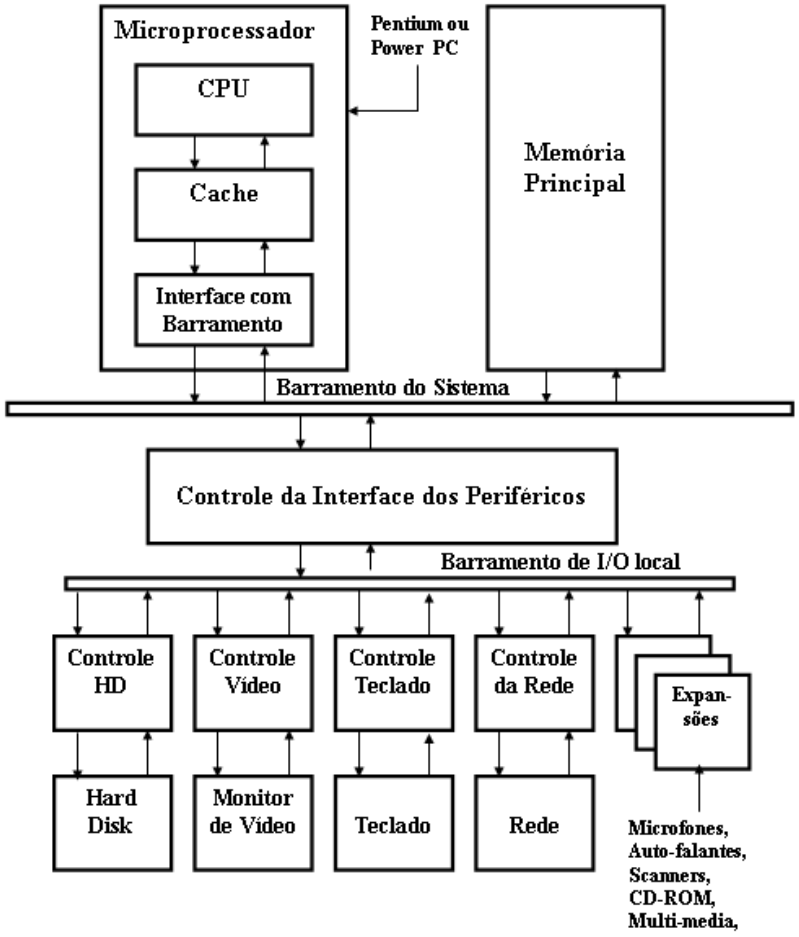
That is, given these assumptions about the MTTF of components, something in a rack fails on average every 3 weeks. About 70% of the failures would be the disks, and about 20% would be fans or power supplies.

Barramentos

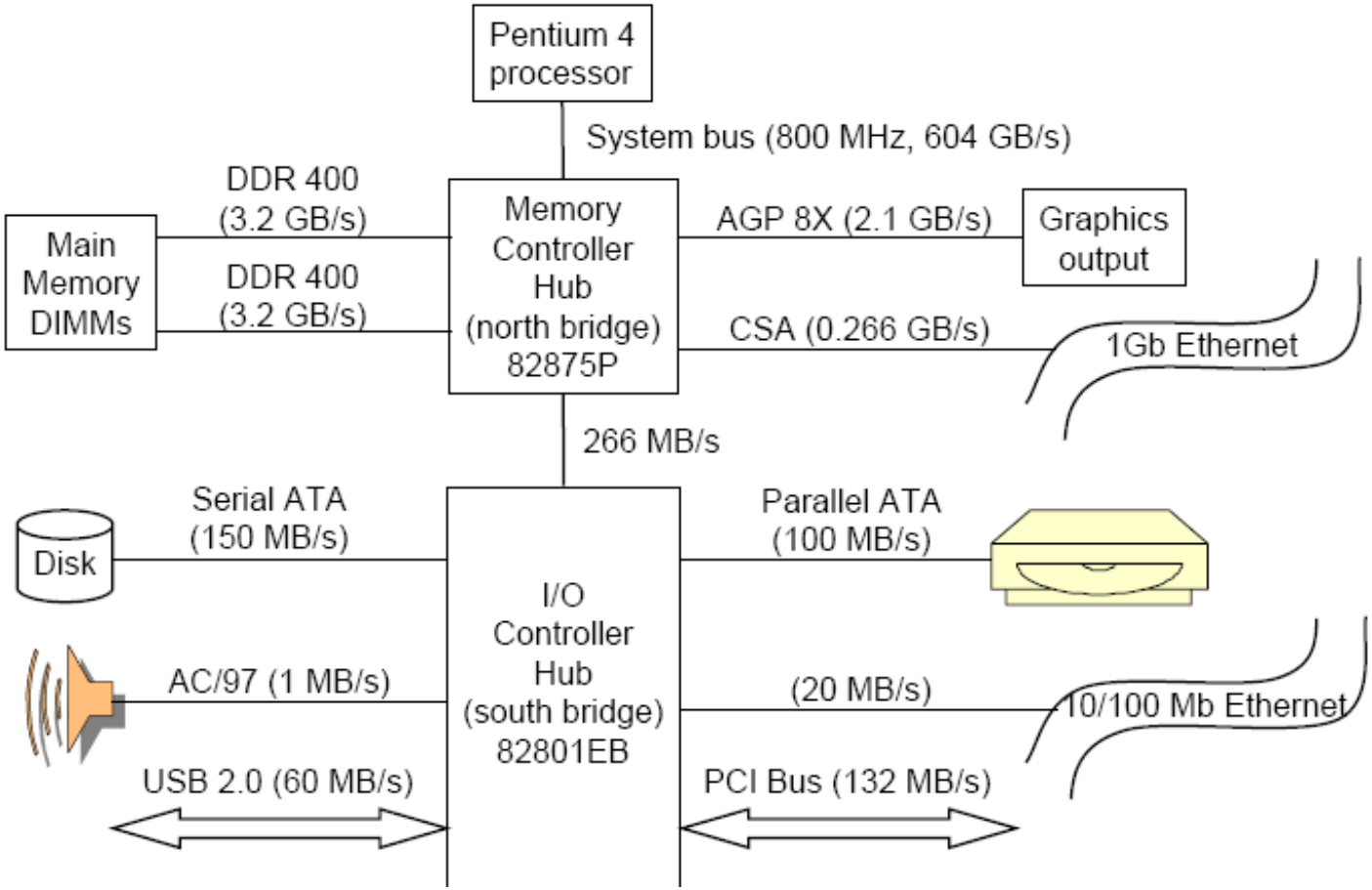
CES-25 – Arquiteturas para Alto Desmpenho
Prof. Paulo André Castro
pauloac@ita.br
Sala 110 – Prédio da Computação
www.comp.ita.br/~pauloac
IEC - ITA

Barramentos





Barramentos no Pentium 4

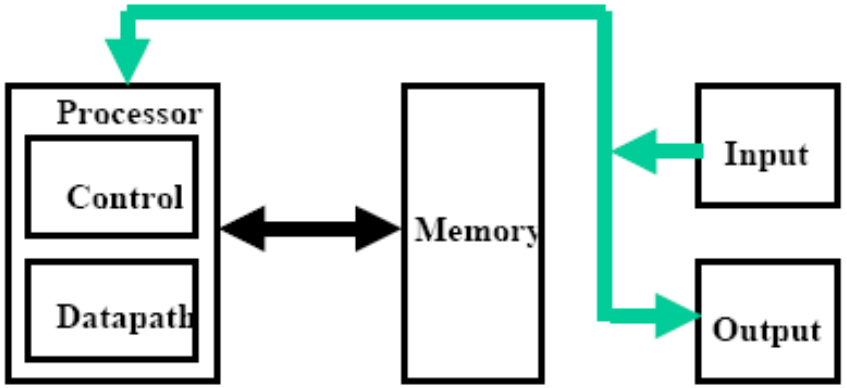


Exemplo de Dispositvos de IO e Tx. De Transferência

Device	Behavior	Partner	Data Rate (Mb/s)
Keyboard	Input	Human	0.0001
Mouse	Input	Human	0.0038
Voice Input	Input	Human	0.2640
Voice Output	Output	Human	0.2640
Scanner	Input	Machine	3.2
Laser Printer	Output	Human	3.2
Sound Input	Input	Machine	3.0
Sound Output	Output	Human	8.0
Wireless LAN	Input or Output	Machine	11 – 54
Magnetic Tape	Storage	Machine	32
Optical Disk	Storage	Machine	80
Network/LAN	Input or Output	Machine	100 – 1,000
Magnetic Disk	Storage	Machine	240 – 2,560
Graphics Display	Output	Human	800 – 8,000

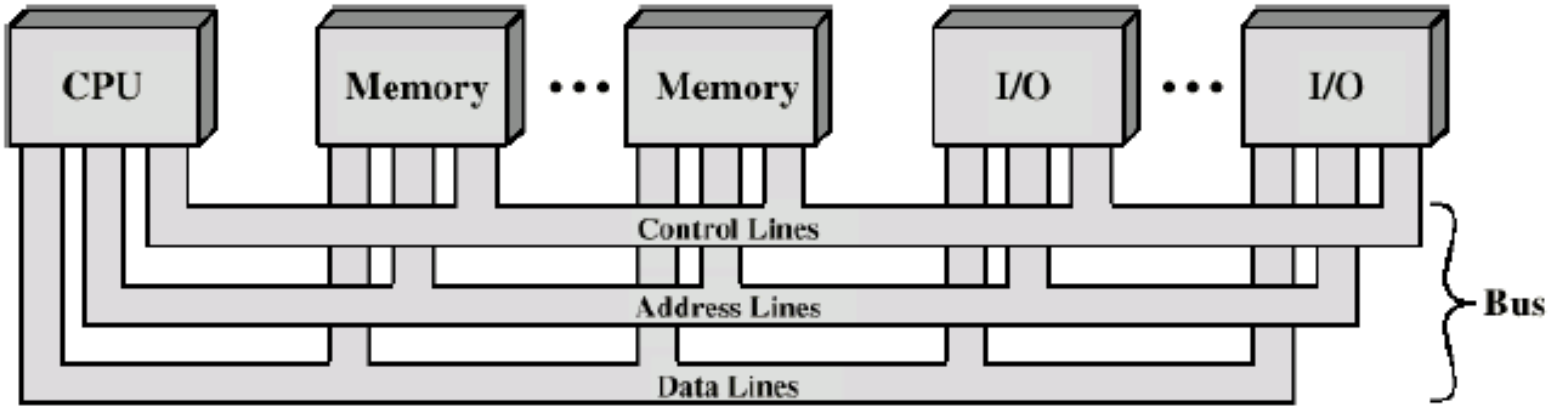
Um Barramento (Bus) é

- A ***shared communication link*** for conveying addresses, data, and control signals
- A single set of wires used to connect multiple subsystems



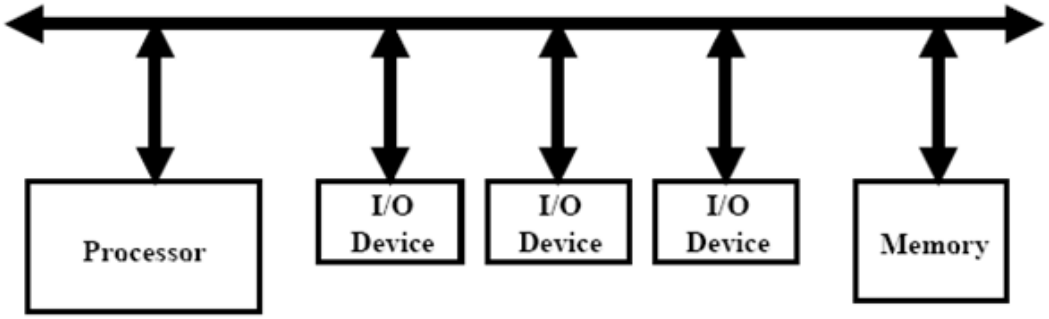
- A fundamental tool for composing large, complex systems
  - systematic means of abstraction

Barramentos



- Barramento de Dados
  - Transporta dados (ou instruções) não há diferença neste nível
  - Geralmente bidirecional
  - Largura é determinante para o desempenho
- Barramento de Controle
  - Sinais de Controle (ler/gravar)
  - Sinais de interrupção
  - Sinais de clock
- Barramento de Endereços
  - Identifica fonte ou origem de um fluxo de dados
  - Largura identifica a capacidade máxima de endereçamento

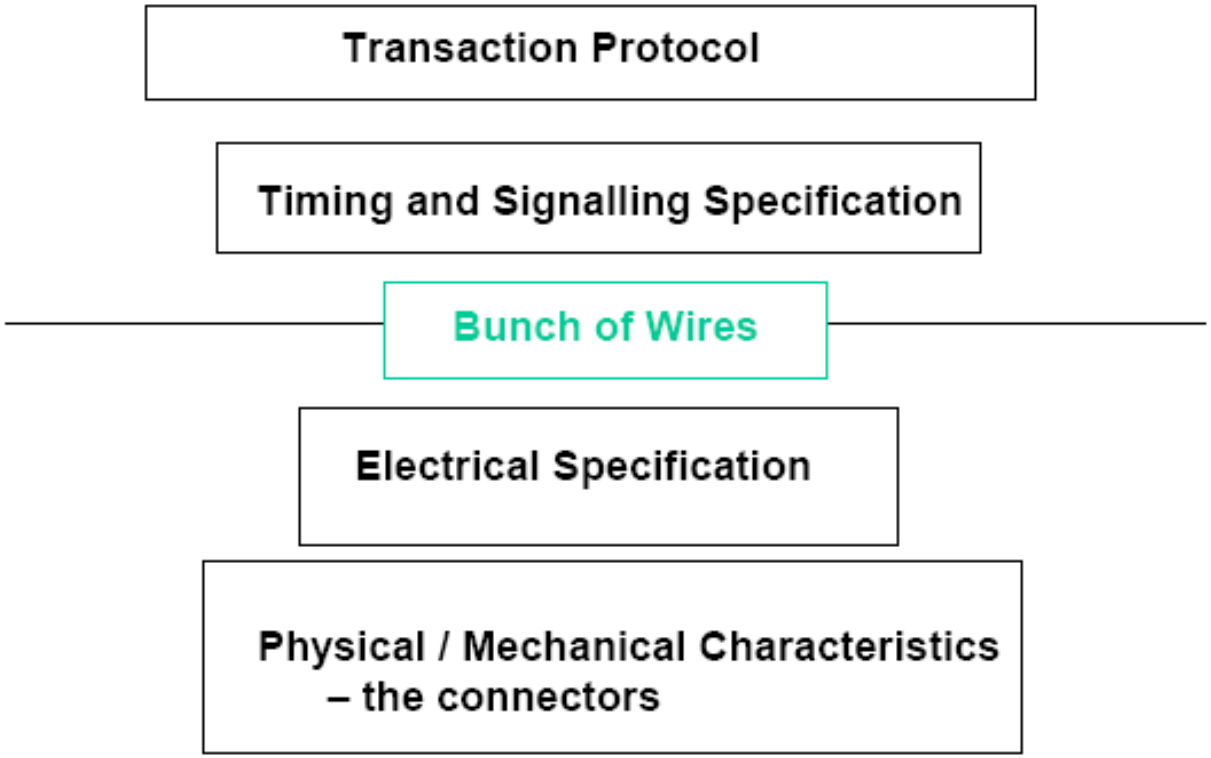
Desvantagens e Vantagens de Barramentos



- It creates a communication bottleneck
  - The bandwidth of the bus can limit the maximum I/O throughput
- The maximum bus speed is largely limited by:
  - The length of the bus
  - The number of devices on the bus
  - The need to support a range of devices with:
    - Widely varying latencies
    - Widely varying data transfer rates

- Versatility:
  - New devices can be added easily
  - Peripherals can be moved between computer systems that use the same bus standard
- Low Cost:
  - A single set of wires is shared in multiple ways
- Manage complexity by partitioning the design into multiple, separate modules

O que define um Barramento ?



Projeto de Barramentos

- A velocidade e a largura de banda são influenciados por 4 fatores principais:
  - Largura do Barramento
  - Esquema de Clock do Barramento
  - Método de Arbitragem
  - Operação

Largura do Barramento

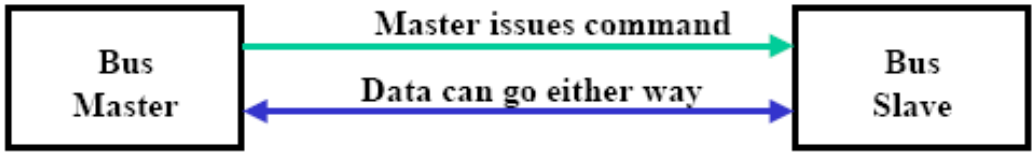
- O número de linhas de endereço determina o tamanho da memória endereçável
- Quanto maior o número de linhas, mais fios, conectores maiores. Logo, o hardware torna-se mais caro.
  - 8088 – 20 linhas de endereço, 80286 + 4 linhas, 80386 +8 linhas
- A tendência é um crescimento das larguras dos barramentos para aumentar a capacidade dos barramentos, mas isto cria problemas de conexão física
- Muitas vezes projetistas fazem multiplexação de dados e endereços em diferentes fases (ou em tempo) para reduzir o número de linhas. Mas com isto também se reduz o desempenho do barramento.

Bus Clocking:
Synchronous and Asynchronous Bus

- **Synchronous Bus:**
  - Includes a clock in the control lines
  - A fixed protocol for communication that is relative to the clock
  - Advantage: involves very little logic and can run very fast
  - Disadvantages:
    - Every device on the bus must run at the same clock rate
    - To avoid clock skew, they cannot be long if they are fast

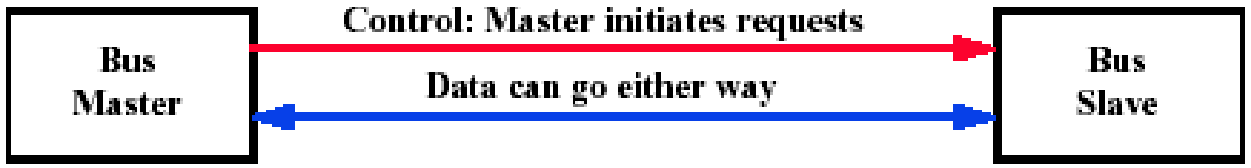
- **Asynchronous Bus:**
  - It is not clocked
  - It can accommodate a wide range of devices
  - It can be lengthened without worrying about clock skew
  - It requires a handshaking protocol

Mestre (Master) e Escravo(Slave)



- A bus transaction includes two parts:
  - Issuing the command (and address) – request
  - Transferring the data – action
- Master is the device that starts the bus transaction by:
  - issuing the command (and address)
- Slave is the device that responds to the address by:
  - Sending data to the master if the master asks for data
  - Receiving data from the master if the master wants to send data

Obtaining Access to the Bus

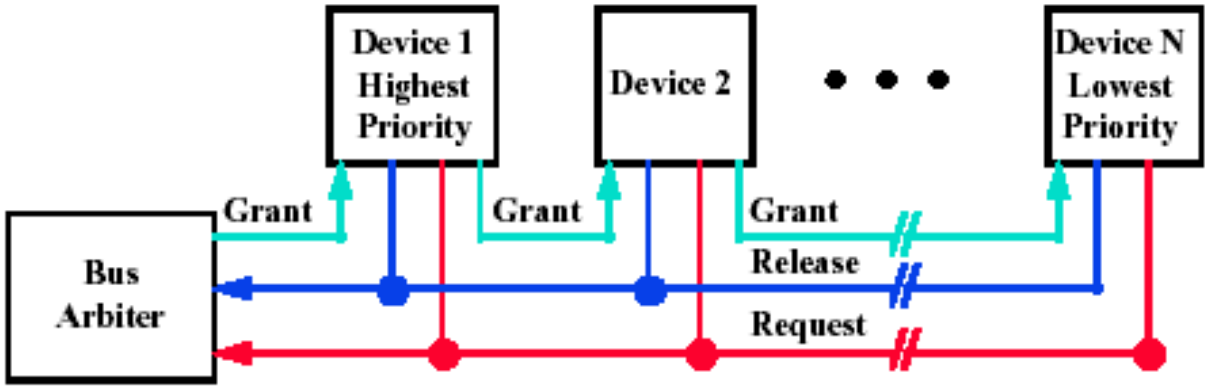


- One of the most important issues in bus design:
  - How is the bus reserved by a devices that wishes to use it?
- Chaos is avoided by a master-slave arrangement:
  - Only the bus master can control access to the bus:
    It initiates and controls all bus requests
  - A slave responds to read and write requests
- The simplest system:
  - Processor is the only bus master
  - All bus requests must be controlled by the processor
  - Major drawback: the processor is involved in every transaction

Múltiplos Mestres de Barramento: Arbitragem

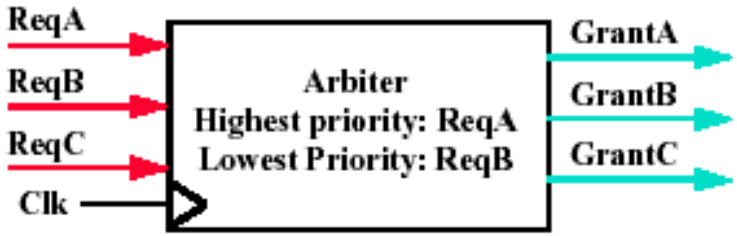
- Com múltiplos possíveis mestres de barramento é necessário definir um meio de garantir que apenas um dispositivo será selecionado como mestre.
- O método deve balancear:
  - Prioridade entre dispositivos
  - Justiça: mesmo o dispositivo de prioridade mais baixa deve operar
- Quatro possíveis Classes de Arbitragem
  - Arbitragem distribuída por auto-seleção: Cada dispositivo coloca o próprio código
  - Arbitragem distribuída por detecção de colisão: exemplo Ethernet
  - Daisy Chain: Autorização dada em seqüência...
  - Arbitragem Centralizada: Autorização dada por órgão central...

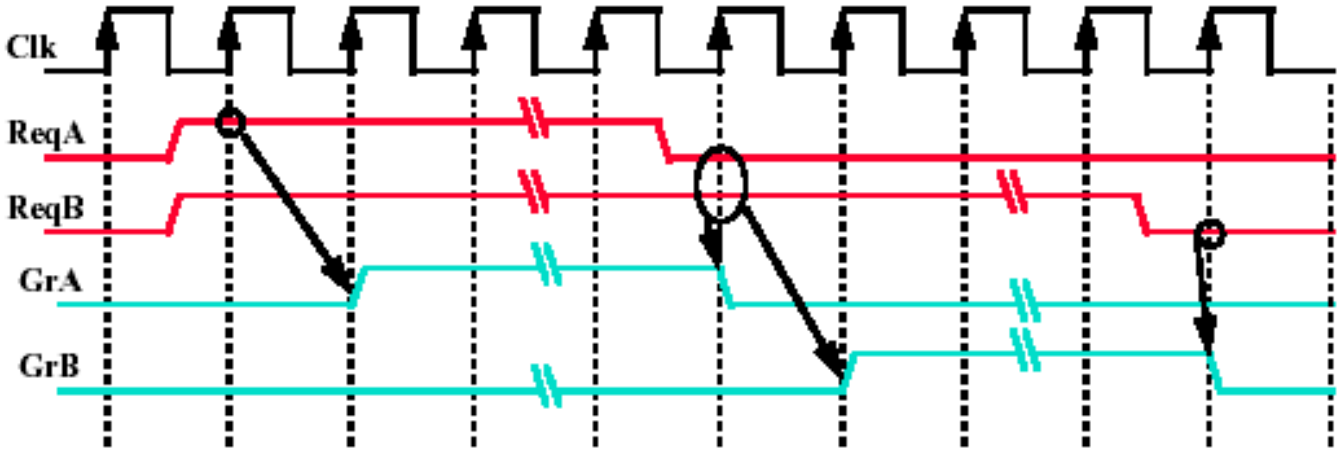
The Daisy Chain Bus Arbitrations Scheme



- Advantage: simple
- Disadvantages:
  - Cannot assure fairness:
    A low-priority device may be locked out indefinitely
  - The use of the daisy chain grant signal also limits the bus speed

Centralized Arbitration with a Bus Arbiter





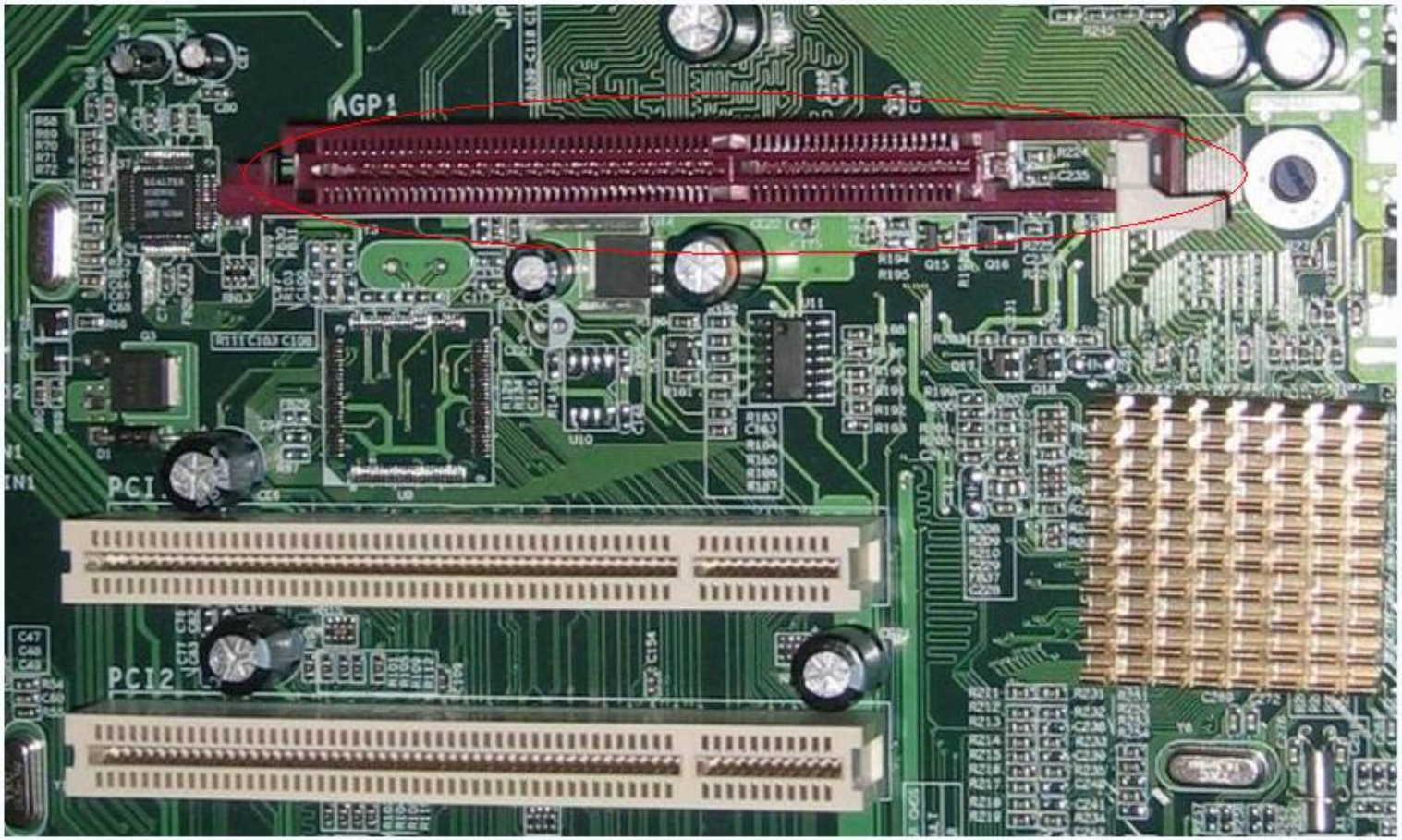
Exemplos de Barramentos

- PCI, PCI Express
- SCSI
- USB, USB 2.0, 3.0
- IEEE 1394 (Fireware)
- ISA, EISA
- VESA
- MIL STD 1553
- Arinc 429
- AGP: Uma porta para o vídeo
- Infiniband : Tipicamente, usado em clusters (racks)

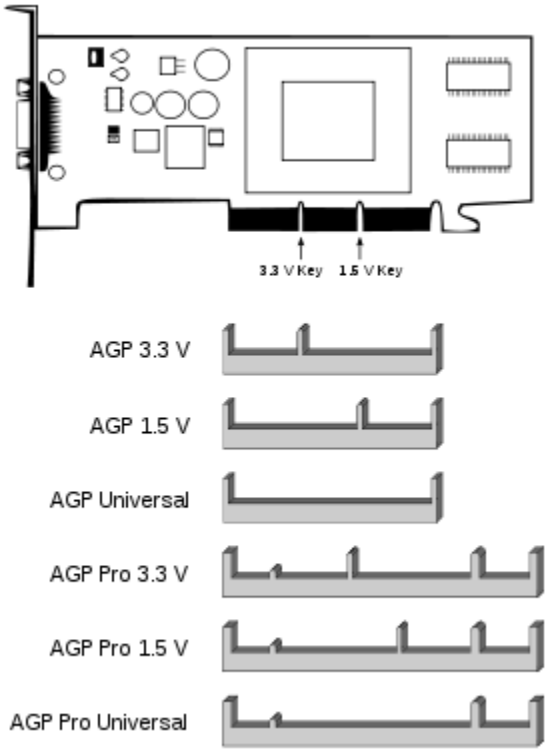
Vídeo
AGP: Uma porta para vídeo

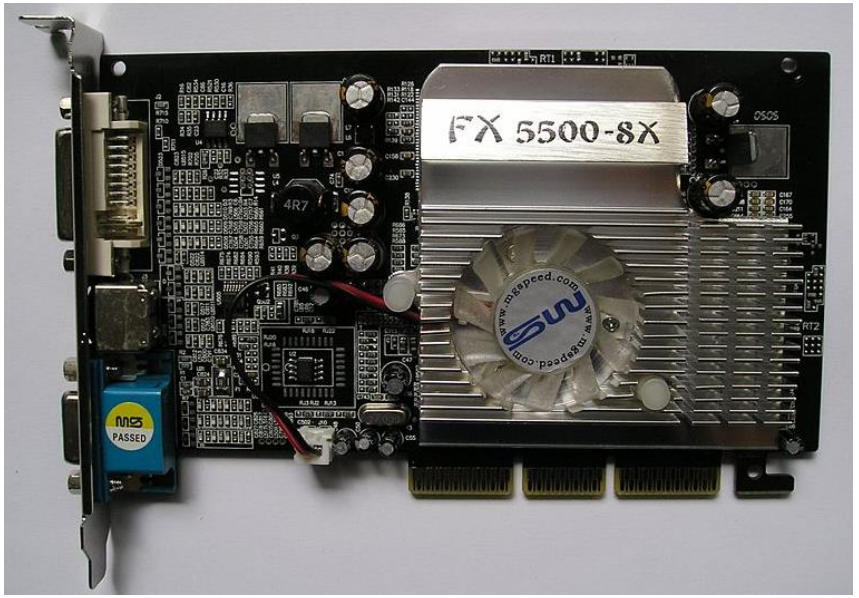
- Solução encontrada para aproximar a ligação da placa de vídeo do processador por causa de sua importância nas aplicações atuais (Interfaces gráficas, jogos, computação gráfica)
- Representa uma "promoção" para a placa de vídeo (sobe na hierarquia)
- É uma porta e não um barramento, ou seja, uma conexão dedicada a um único dispositivo, sem disputa
- Pode ser visto quase como uma ligação direta ao processador (quase porque normalmente esta ligada na ponte do processador)
- Introduzida no Pentium II, chipset 440LX
- Slot parecido com o PCI, um pouco menor e mais alto
- Largura de 32 bits com 66 MHz tendo uma vazão de 254.3 MBytes/s
- Melhorias no padrão indicadas por 2x (transmissão de 2 dados por ciclo), 4x (transmissão de 4 dados por ciclo) resultando em aumento de vazão
- Cuidado: não adianta acelerar só a conexão entre bridge e placa de vídeo se ambos os extremos não suportarem a mesma vazão (o barramento interno de dados do processador e a placa de vídeo ligada no AGP)

Conector AGP

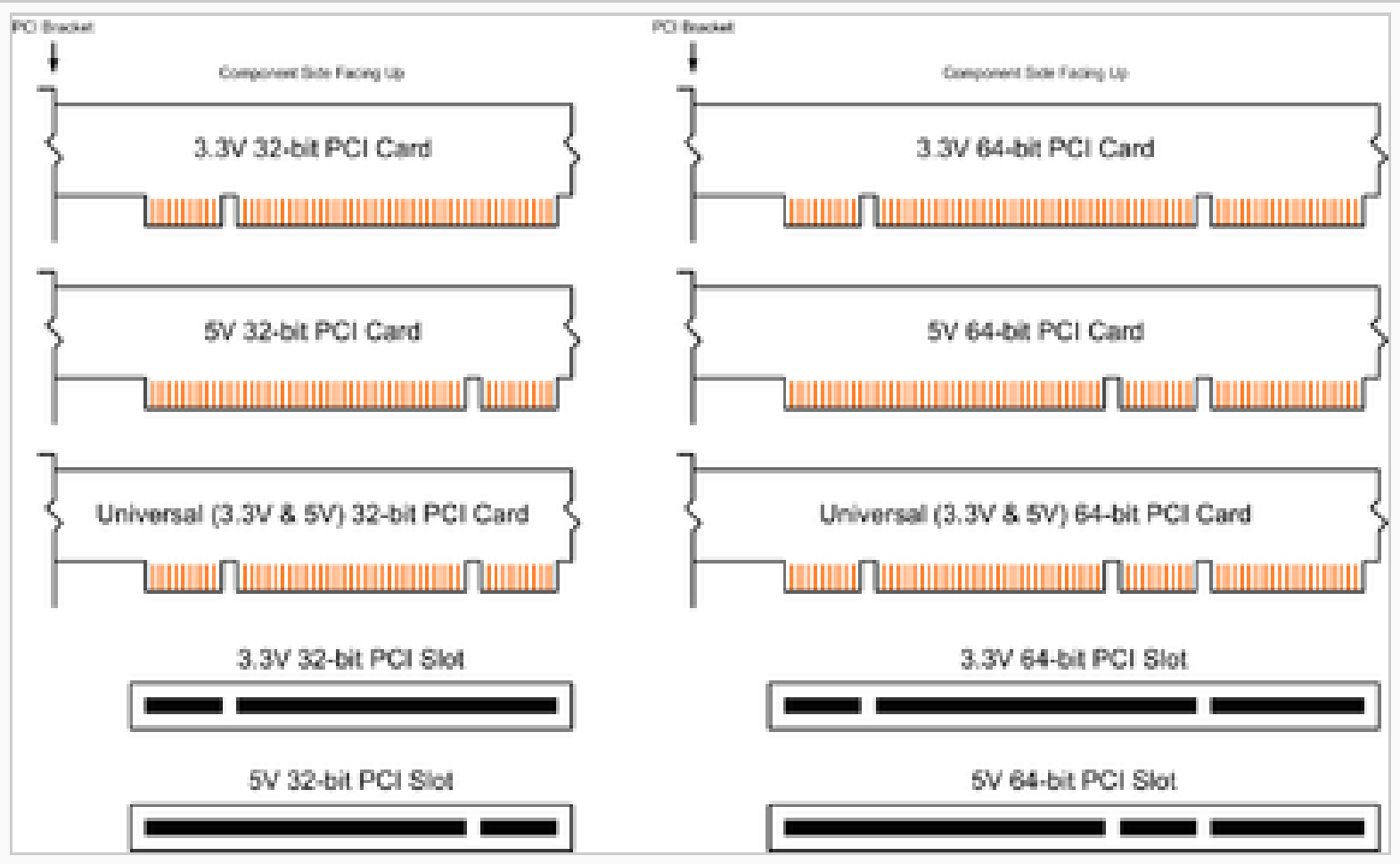


Tipos de Conectores AGP e Placa

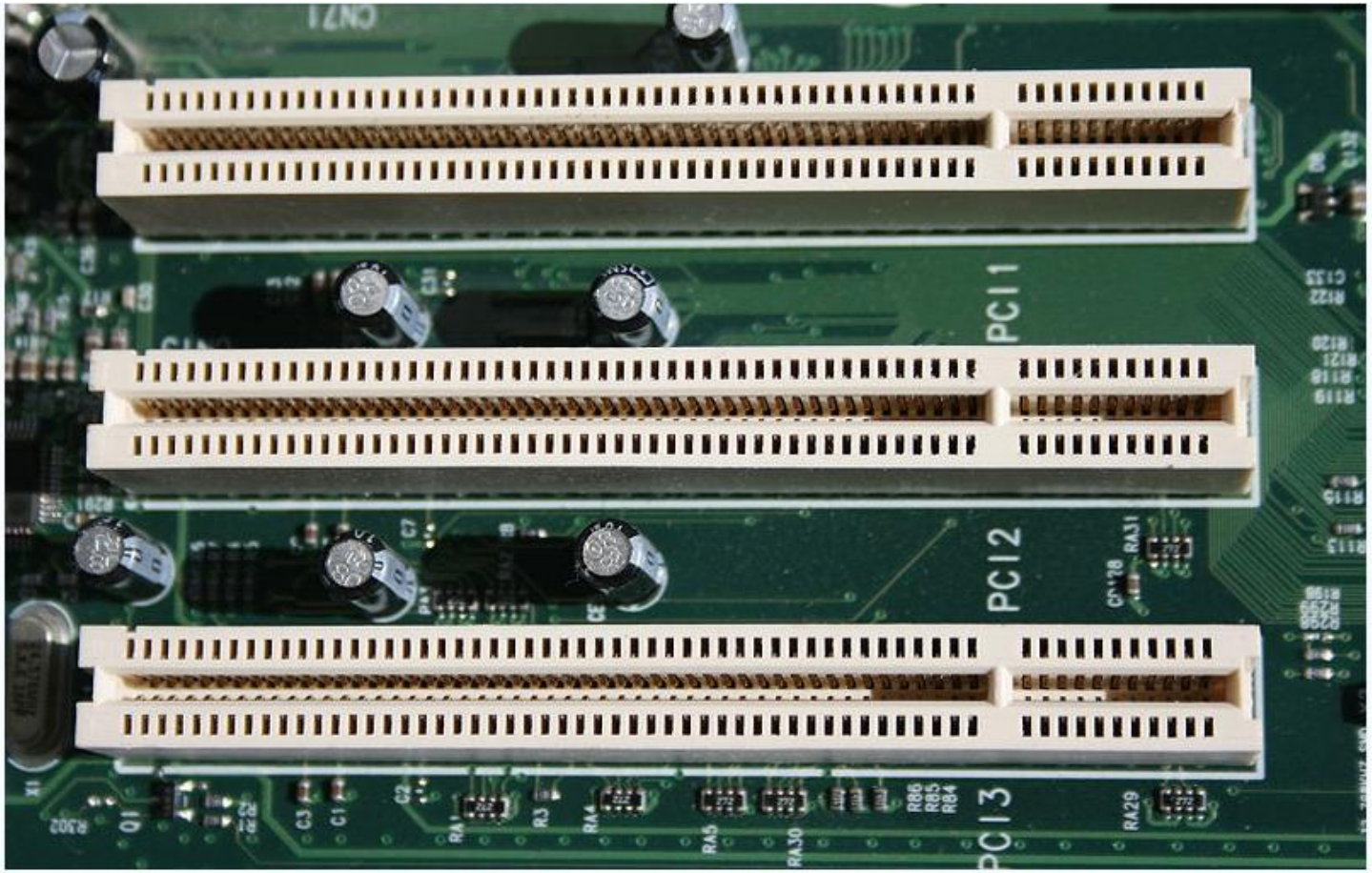




Barramento PCI

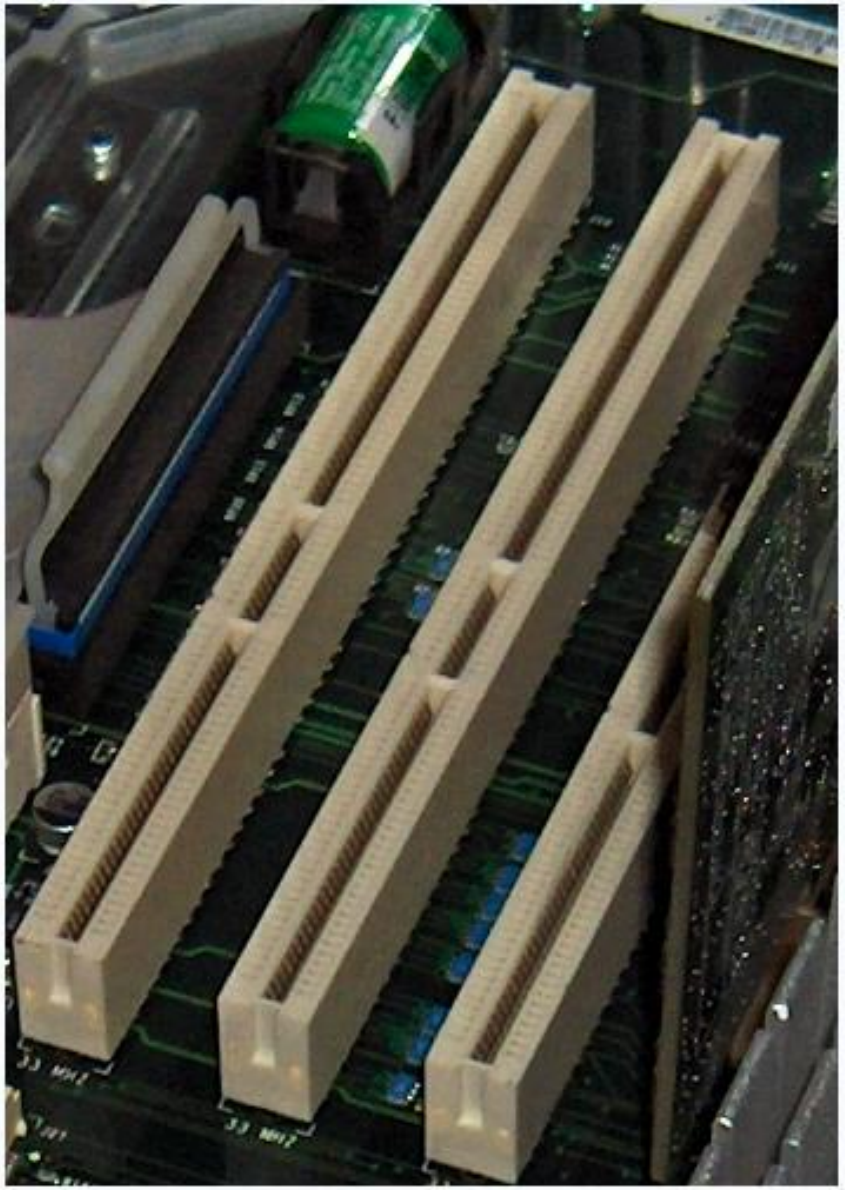


Conectores PCI 32 bits 5V em uma Placa Mãe

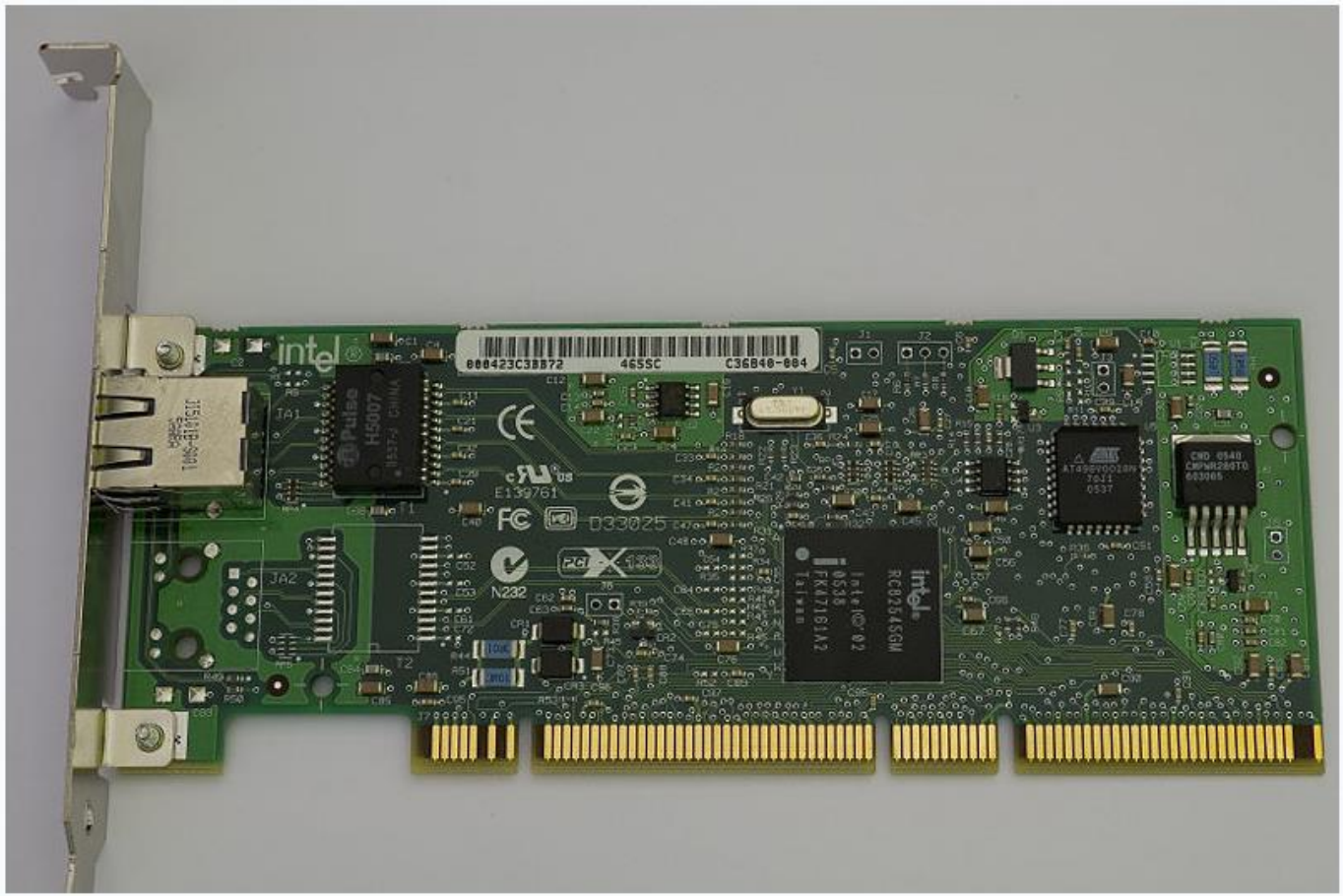


Exemplo de Placa PCI – 32 bits
Adaptador SCSI

Conectores PCI 64 bits 5V em uma Placa Mãe



Exemplo de Placa PCI – 64 bits
Universal Placa de Rede Ethernet



Barramentos Antigos...ISA

- ISA, (Industry Standard Architecture)
  - palavra de 8 bits
  - 62 pinos,
  - taxa de **1.2 MB/s)**

Barramentos Antigos...VESA

- VESA Local Bus (VLB)
  - Palavra de 32 bits, 112 pinos
  - Taxa de transferência:133 MB/s
  - Slot é uma extensão do ISA

Universal Serial Bus

- Versões USB:
  - *USB 1.0*: 1996.
    Taxas de transferências de 1.*5 Mbit/s* (*Low-Speed)* até *12 Mbit/s* (*Full-Speed*).
  - *USB 1.1*: 1998.
    Correção de alguns problemas (bugs) da primeira versão. Esta foi a primeira versão USB a ser amplamente utilizada.
  - USB 2.0: 2000.
    - Adicionou um novo modo “High speed” que permite taxas de até 480Mbps
  - *USB 3.0*. 2008.
    - Taxa de transferência de até 5Gbps

Barramentos Aviônicos

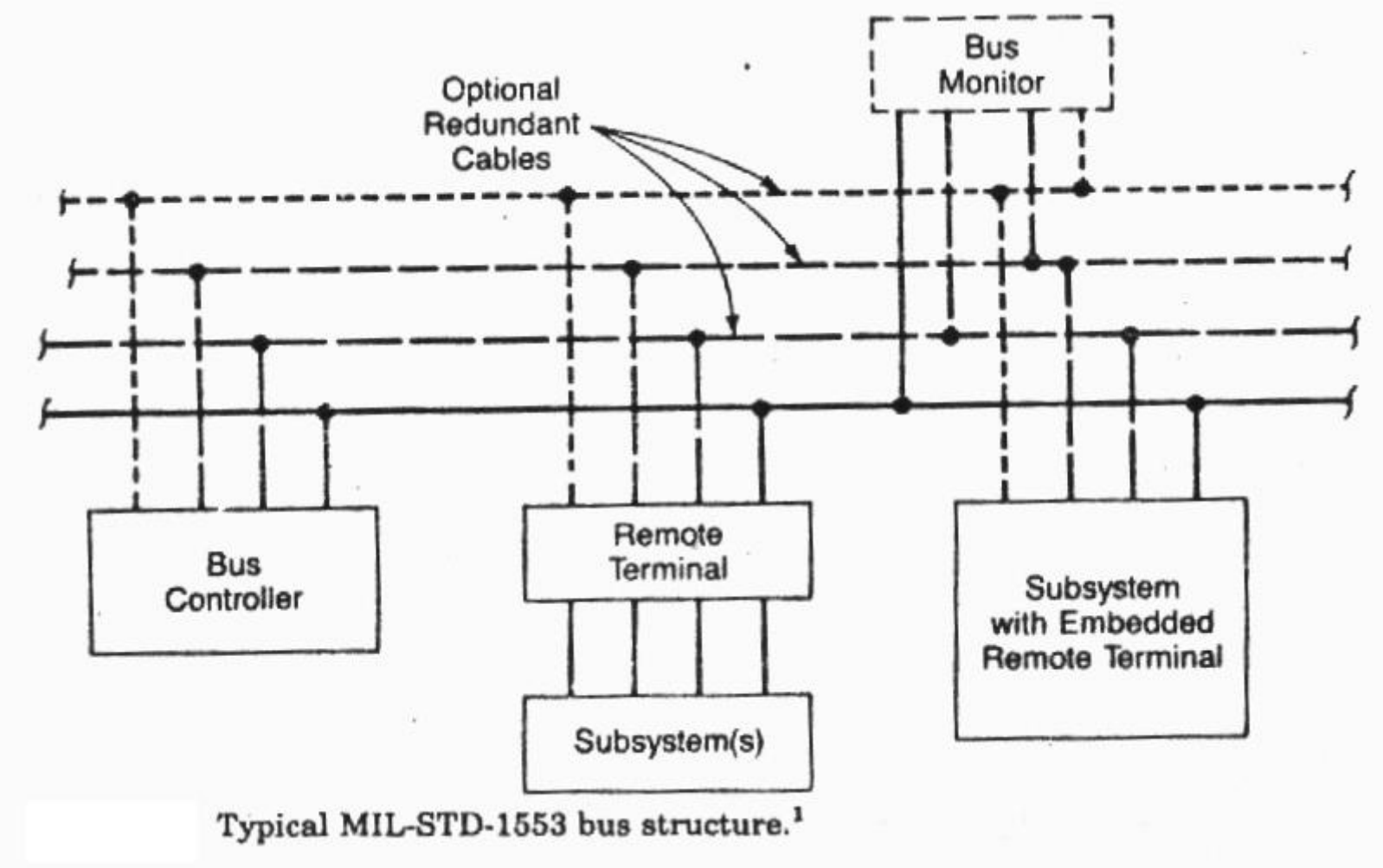
- MIL-STD 1553 (Padrão para aviões militares)

- Uso Civil
  - ARINC 429
  - ARINC 629

MIL-STD 1553

- O padrão 1553 pode ser dividido em três partes:
  - Tipos de terminais: Bus controller, Bus monitor (opcional) e Remote Terminal
  - Protocolo de Barramento: incluindo formatos de mensagens e estrutura
  - Especificação de hardware: tais como impedâncias, frequência de operação, etc.
- O barramento 1553 pode operar com até 1Mbps de taxa de transferência

MIL-STD 1553 – Estrutura de Barramento



Typical MIL-STD-1553 bus structure.[1]

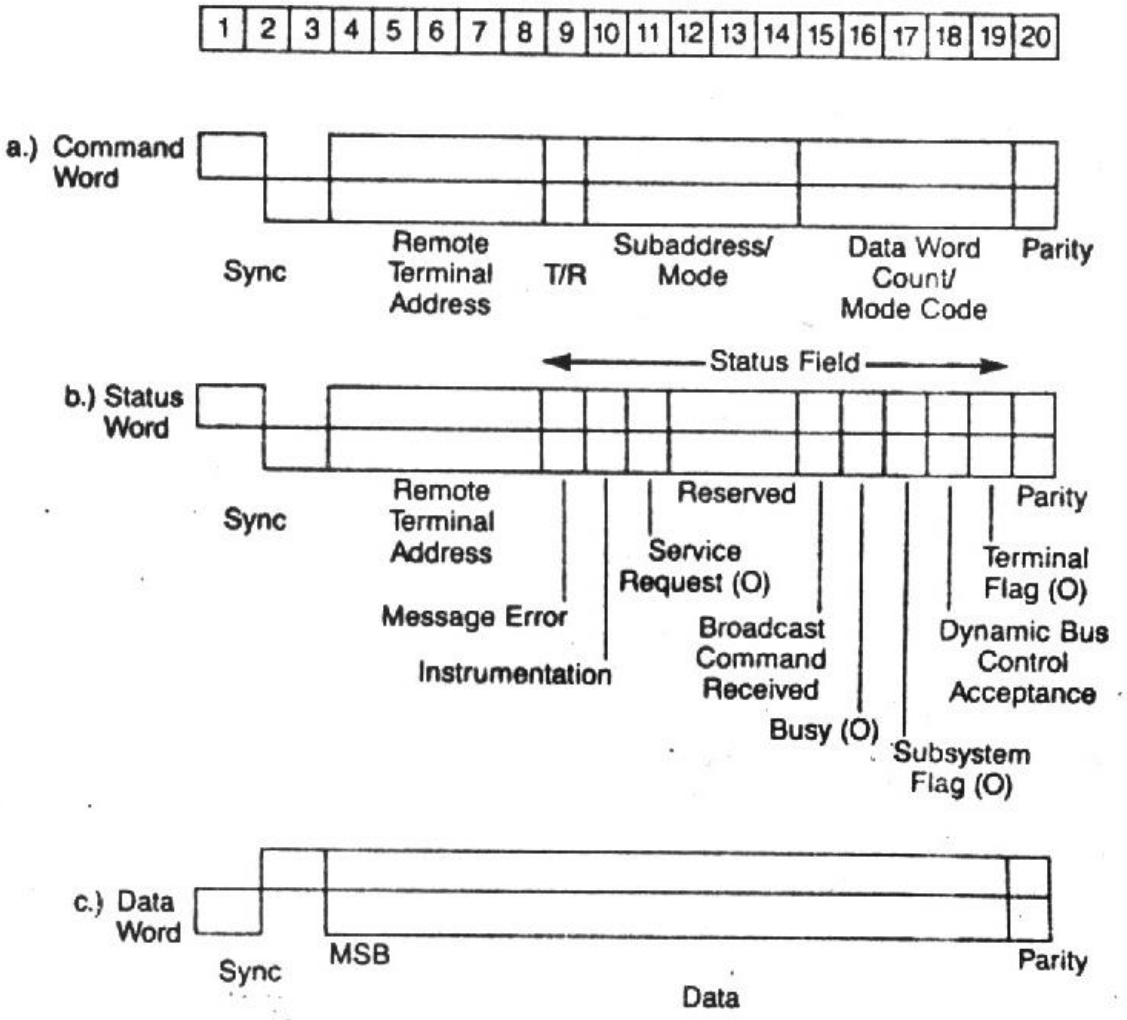
Tipos de Terminais no padrão1553

- Bus Controller: Responsável por todo o fluxo de dados do barramento e inicia todas as transferências de informação. Também monitora o status do sistemas, não confundir com o Bus monitor.

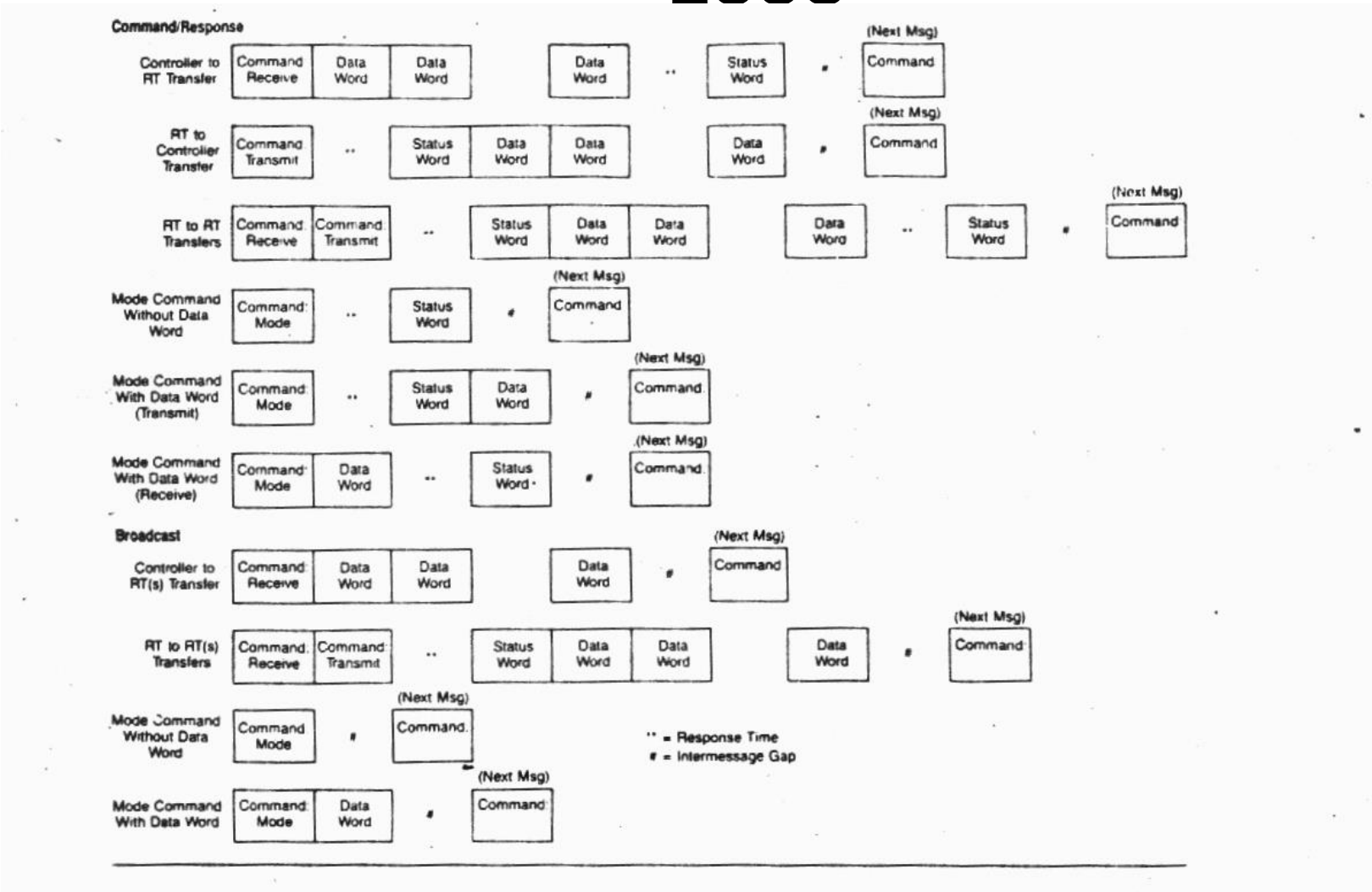
- Bus Monitor: Recebe e armazena tráfego selecionado no barramento. Não responde a nenhum tráfego

- Remote Terminal: São o maior número de unidades de um barramento 1553. Devido a endereçamento de RT utilizar 5 bits nas mensagens, podem existir até 31 RT em um barramento. Um RT pode ser uma unidade separada para ligar um subsistema ou ser parte do subsistema.

Words 1553


1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
a.) Command Word
Sync
Remote Terminal Address
T/R
Subaddress/ Mode
Data Word Count/ Mode Code
Parity
Status Field
b.) Status Word
Sync
Remote Terminal Address
Message Error
Instrumentation
Reserved
Service Request (O)
Broadcast Command Received
Busy (O)
Subsystem Flag (O)
Dynamic Bus Control Acceptance
Terminal Flag (O)
Parity
c.) Data Word
Sync
MSB
Data
Parity

Formatos de Transferência de Dados - 1553

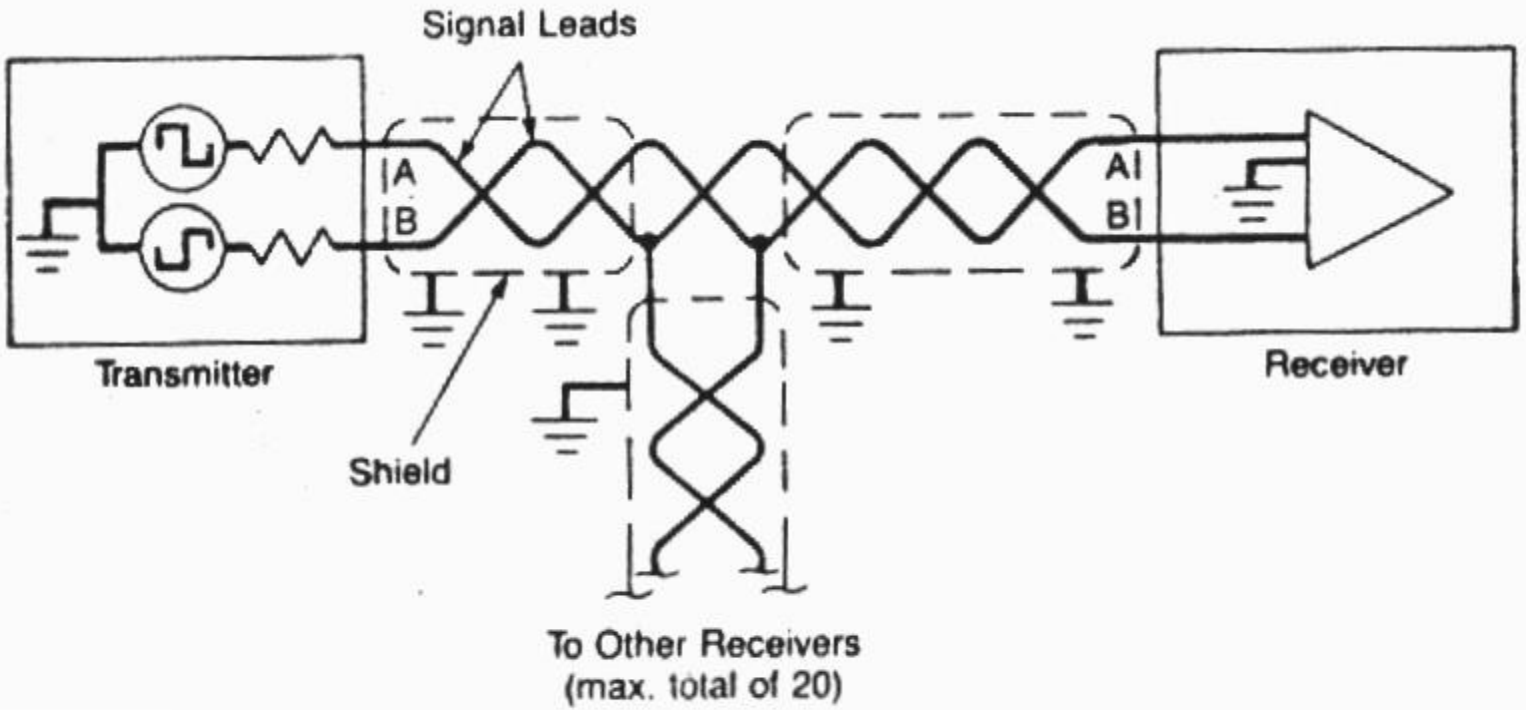


ARINC 429

- “ARINC. Specification 29 Digital Information Transfer System, Mark 33”, 429 as it is commohnly known , is the basis from digital buses in modern civil aircraft” Digital Avionics Systems. P.31.

- ARINC 429 opera com taxas de transferências de 12 a 14.5 or 100kpbs em um barramento simplex

Barramento Arinc 429



Generalized ARINC 429 bus.[8]