

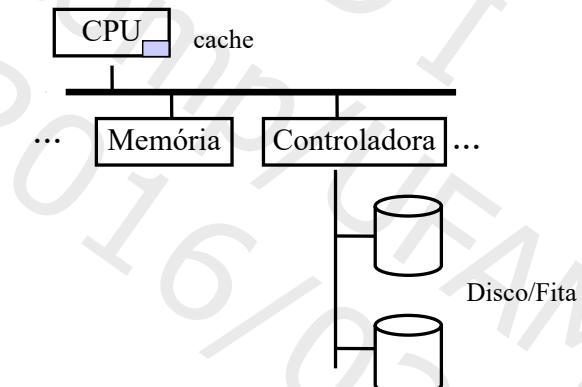
# Bancos de Dados I

Conceitos de Memória Secundária  
Prof. Altigran Soares da Silva



V2018.1

# Hierarquia de Armazenamento



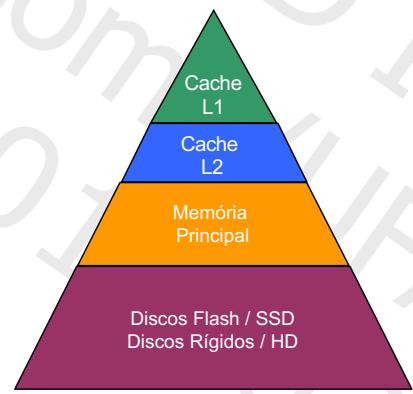
# Meios de Armazenamento

- Cache: dentro ou fora da CPU
  - CPU cada vez mais rápida: vários núcleos, clocks de ~3 GHz hoje.
- Memória Principal
  - US\$ 8 por Gigabyte – reduz a cada ano
  - Volátil – não se mantêm entre paradas do sistema
  - Acesso randômico muito rápido
  - Dados processados pela CPU diretamente
  - Capacidade típica limitada.
    - Muito abaixo do que é necessário em bancos de dados.

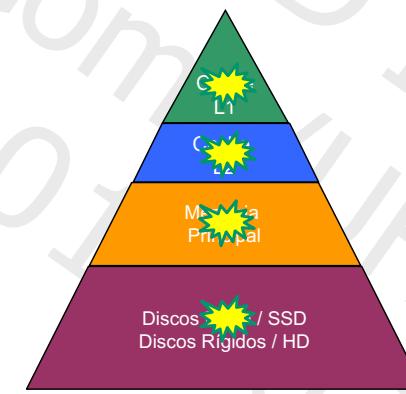
# Meios de Armazenamento

- Disco Magnético
  - Custo de US\$ 0.08 por Gigabyte – reduz a cada ano
  - Não-volátil. Exceto no caso de defeitos no disco
  - Acesso radônico lento
  - CPU não acessa diretamente estes dados. Dados precisam ser transferidos para a memória principal
- Discos Flash/SSD
  - Mais Caro: US\$ 1 por Gigabyte
  - Mais rápido, inclusive para acesso randômico

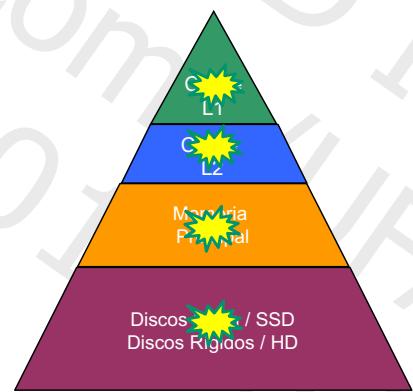
## Pirâmide de Armazenamento



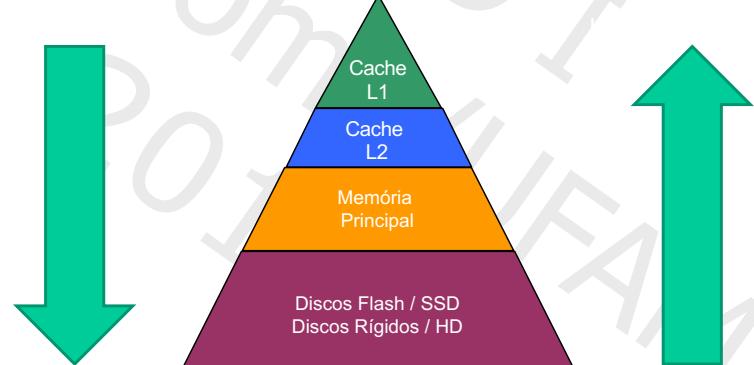
## Pirâmide de Armazenamento - Leitura



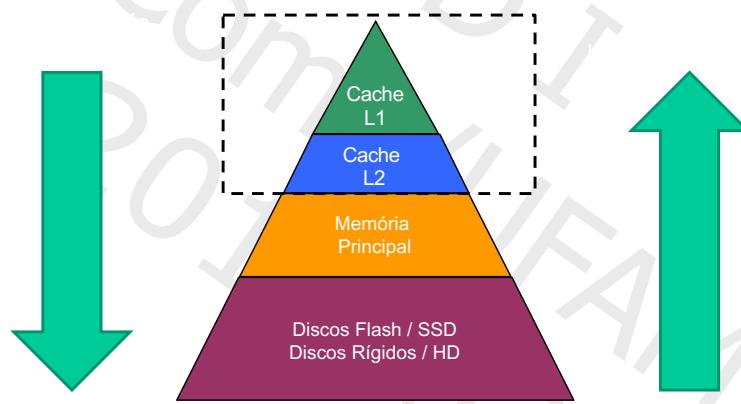
## Pirâmide de Armazenamento - Escrita



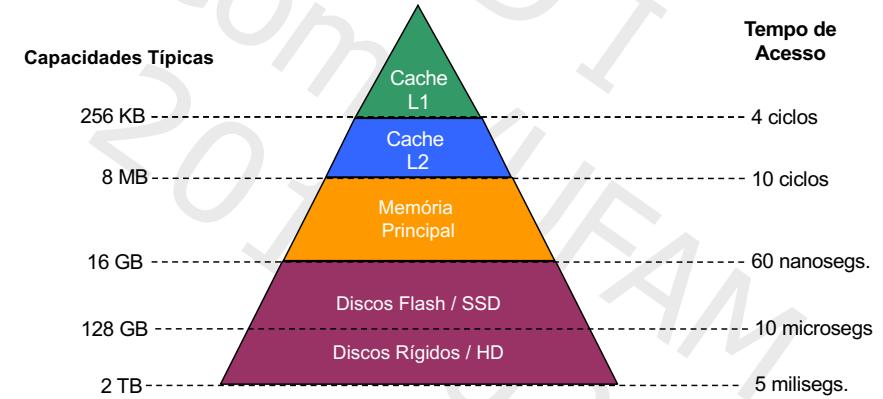
## Pirâmide de Armazenamento



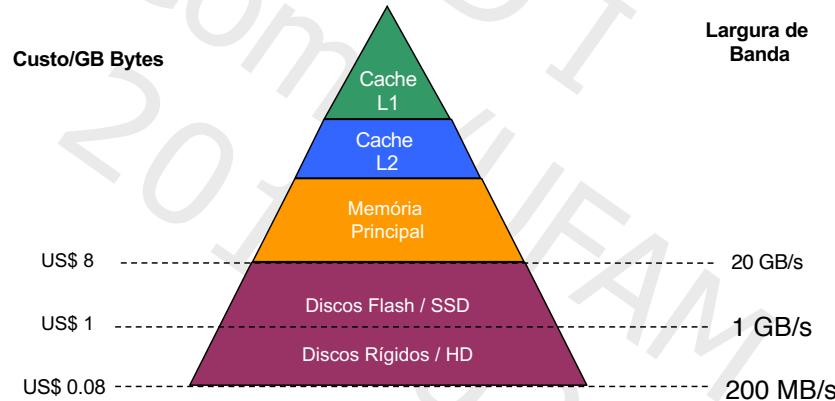
## Pirâmide de Armazenamento



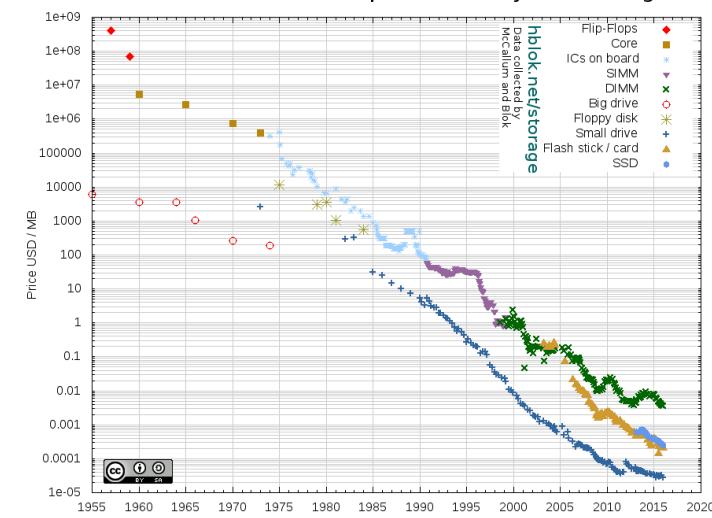
## Pirâmide de Armazenamento



## Pirâmide de Armazenamento

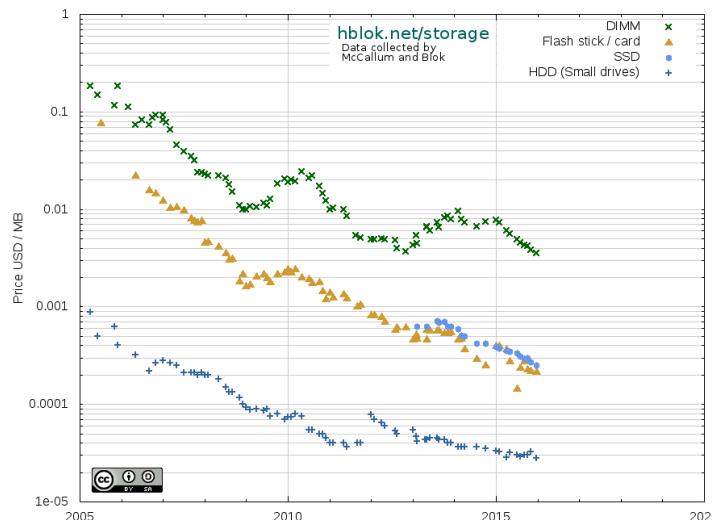


Historical Cost of Computer Memory and Storage



Fonte: <https://hblok.net/blog/posts/2013/02/13/historical-cost-of-computer-memory-and-storage/>

## Historical Cost of Computer Memory and Storage



## Diferença Relativa de Tempos

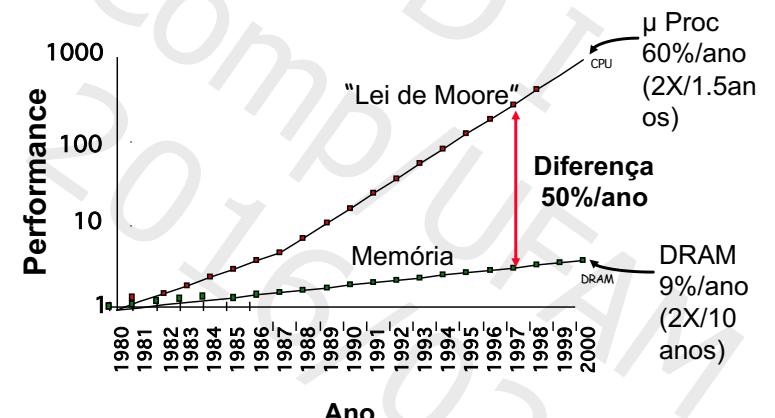
Evento	Real (Aproximado)	Relativo
1 Ciclo de CPU	0.3 ns	1 s
Acesso cache L1	0.9 ns	3 s
Acesso cache L2	2.8 ns	9 s
Acesso cache L3	12.9 ns	43 s
Acesso Memória	120 ns	6 min
Acesso SSD	50-150 µs	2-6 dias
Acesso HDD	1-10 ms	1-12 meses
Internet: SF para NY	40 ms	4 anos
Internet: SF para UK	81 ms	8 anos
Internet: SF para AU	183 ms	19 anos

Fonte: <https://hblok.net/blog/posts/2013/02/13/historical-cost-of-computer-memory-and-storage/>

## Diferença Relativa de Tamanhos



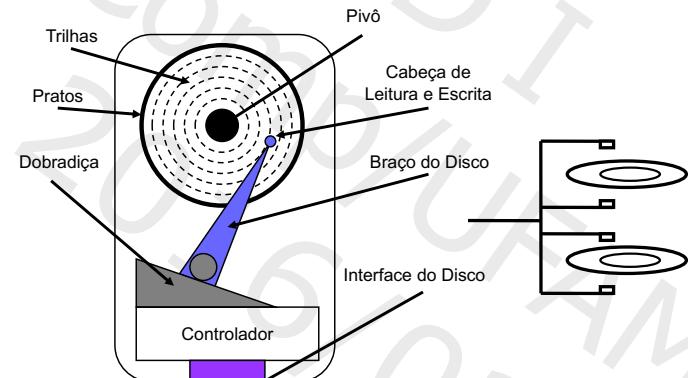
## Lei de Moore e Evolução das Memórias



Fonte: Mohamed Zahrani (NYU) Computer Systems Organization slides

## DISCOS MAGNÉTICOS

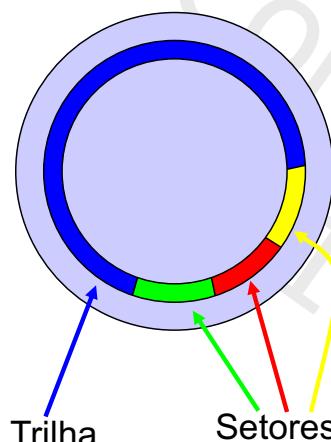
### Discos Magnéticos



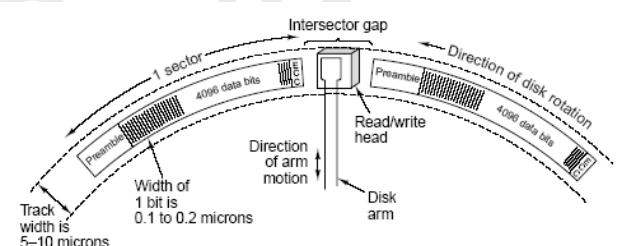
### Discos Magnéticos (2)



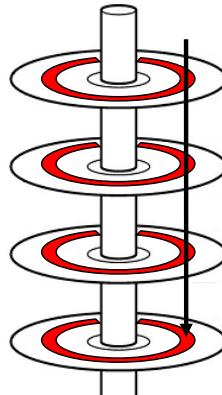
### Discos Magnéticos (3)



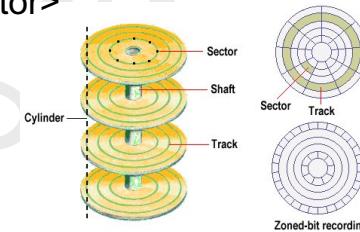
- Cada disco é dividido em regiões anulares chamadas trilhas
  - 500 a 200 trilhas por superfície



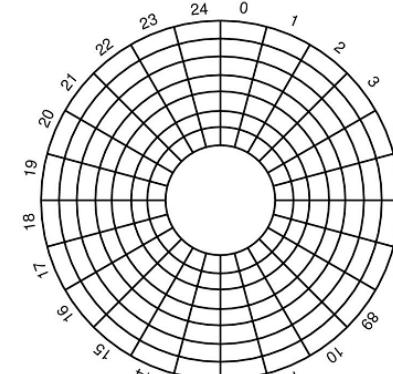
## Discos Magnéticos (4)



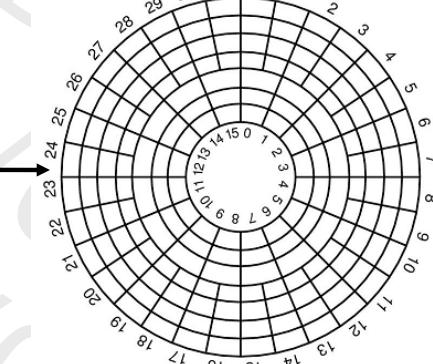
- O conjunto das trilhas acessíveis de uma dada posição do braço formam um cilindro
- O nr. de cilindros é igual ao nr. de trilhas em cada lado dos pratos
- Um local do disco especificado pela tripla <cilindro, cabeça, setor> ou <trilha, cabeça setor>



## Setores por Trilha (2)



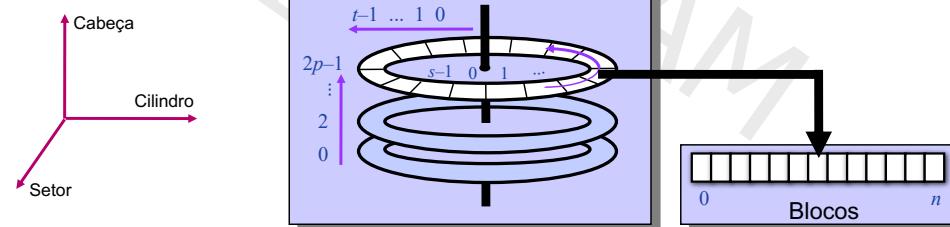
Geometria Virtual



Geometria Física

## Mapeamento Lógico de Endereços (2)

- Os discos são endereçados como um longo arranjo unidimensional onde blocos lógicos são a menor unidade de transferência para a memória principal.
- Os blocos são mapeados para setores do disco de maneira sequêncial.
  - O setor 0 é o primeiro setor da primeira trilha do cilindro mais externo.
  - O mapeamento prossegue pela mesma trilha, depois pelas outras trilhas do mesmo cilindro e dai para os outros cilindros, do mais externo para o mais interno.

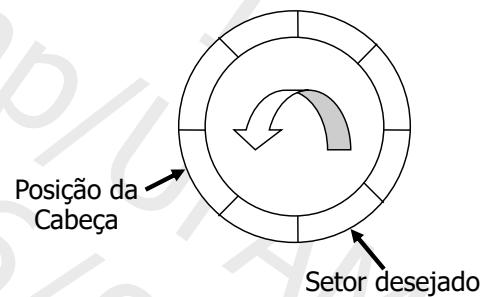


## Tempo de Acesso

- Acesso de leitura ou escrita requer três passos:
  - Seek: busca da trilha; posicionamento do braço na trilha correta
  - Rotação: espera para que o setor desejado seja rotacionado até a cabeça de leitura/escrita
  - Tempo de transferência: Transferência dos bits armazenados no setor que está ao alcance da cabeça.
  - Bloco: Unidade de Transferência
- **Tempo de acesso =**
  - Tempo de Seek + Latência Rotacional + Tempo de Transferência+ Outros

## Latência Rotacional – Valores Típicos

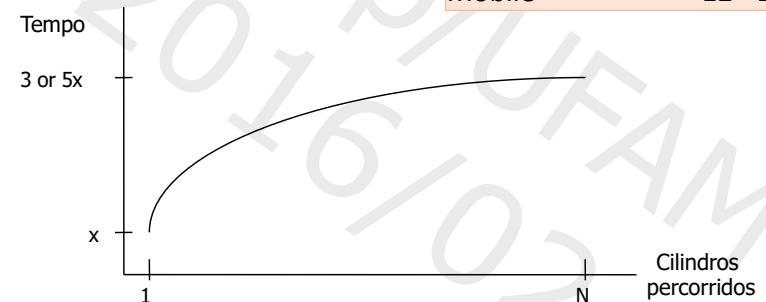
Giro (RPM)	LRM (MS)
4.200	7,14
5.400	5,56
7.200	4,17
10.000	3,00
15.000	2,00



## Tempo de Seek

- Se considerarmos a localidade de referência temos somente 25% a 33% deste tempo

Aplicação	Seek Médio
Servidores Hi-End	4 ms
Desktop	9 ms
Mobile	12 - 15 ms



## Tempo de Transferência

- Valores típicos: 2 a 12 MB por segundo
- Depende de :
  - Tamanho a transferir, usualmente um setor
  - Velocidade de rotação
  - Densidade de gravação: bits por polegada na trilha
  - Diâmetro : 2.5 a 5.25 polegadas

## Outros

- Tempo que leva a CPU para solicitar a operação de E/S
- Concorrência na controladora
- Concorrência para barramento e memória
- Geralmente desprezíveis em comparação de outros tempos

## Exemplos de HDD

Parâmetro	Seagate FC Ultra 160 SCSI	Seagate Cheetah ST373405LC	SEAGATE ST500LX012 Ultra Mobile	WD Archive 8T	SEAGATE Barracuda 7200	WD Blue Mobile WD10SPCX
Capacidade	73 GB	73 GB	500 GB	8 TB	2 TB	1 TB
Pratos/Cabeças	12/24	8/16	1/2	12/06	06/03	-
Trilhas (Média)	14100	776	-	-	-	-
Densidade (Bytes/Setor)	512	-	-	512	-	-
Giro (RPM)	10.000	10.000	5.400	5.900	7.200	5.400
LRM (ms)	2,99	2,99	5,56	5,5	4,17	5,5
Seek (ms)	6	9,4	13	-	8,5	-
Transferência (Mb/S)	160	85	100	190	156	140
Cache (MB)	0.8	4	-	128	64	16

## Acesso Sequencial X Randômico

- Acesso Randômico:
  - Necessita do posicionamento do braço, cabeças, etc.
- Acesso Sequencial
  - Posicionar no “próximo bloco”
    - Saltar o espaço não usado (gap) entre os setores
    - De tempos em tempos: passar para a próxima trilha e depois para o próximo cilindro
- Tempo de acesso sequencial  $\approx$  Tempo de Transferência
- Regra básica:
  - Acesso randômico: Caro
  - Acesso sequencial: Muito mais barato
- Exemplo: para transferir blocos de 1 KB
  - Randômico: ~20 ms.
  - Sequencial: ~1 ms.



## Acesso Sequencial X Randômico

- Acesso Randômico:
  - Necessita do posicionamento do braço, cabeças, etc.
- Acesso Sequencial
  - Posicionar no “próximo bloco”
    - Saltar o espaço não usado (gap) entre os setores
    - De tempos em tempos: passar para a próxima trilha e depois para o próximo cilindro
- Tempo de acesso sequencial  $\approx$  Tempo de Transferência
- Regra básica:
  - Acesso randômico: Caro
  - Acesso sequencial: Muito mais barato
- Exemplo: para transferir blocos de 1 KB
  - Randômico: ~ 20 ms.
  - Sequencial: ~ 1 ms.

## Acesso Sequencial x Randômico

- BD de 1 TB com registros de 100 bytes
  - Vamos atualizar 1% dos registros
- Cenário 1: acesso randômico
  - Cada atualização leva ~30 ms (seek, read, write)
  - 108 atualizações = ~35 dias
- Cenário 2: re-escrever todos os registros
  - Assuma 100 MB/s de taxa de transferência
  - Tempo = 5,6 horas !
- Evitar acessos randômico !

## Você sabia ?

- O primeiro HD a alcançar mais de 1GB foi apresentado pela IBM em 1980



IBM Disk 350, 5MB – Anos 50

- Foram necessários
  - 51 anos para alcançar 1 TB
  - 2 anos para alcançar 2 TB

## Acessos para Escrita

- A não ser que se queira verificar o que foi escrito, o custo dos acessos para escrita é similar ao custo de leitura
- Passos para escrita de um bloco
  - (a) Ler o bloco
  - (b) Modificar em memória
  - (c) Escrever o bloco
  - (d) Verificar - Opcional

## ARMAZENAMENTO FLASH

# Armazenamento Flash

- Dispositivos semi-condutores para armazenamento não-volátil
- Não têm componentes mecânicos
- 100 a 1000 vezes mais rápidos que os discos magnéticos (HDD)
- Menores, consomem menos energia, são mais robustos



# Armazenamento Flash - Degradiação

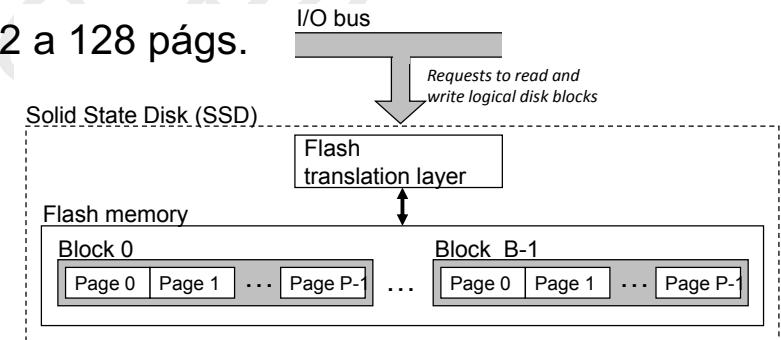
- Memória flash permitem somente um número finito de modificações (escrita/exclusão).
- Tipicamente 10.000 a 1.000.000 de ciclos de limpeza.
- Parcialmente compensado por software: Bloco são remapeados dinamicamente para que as escritas sejam mais uniformemente distribuídas
- Na prática, a degradação levará centenas de anos para acontecer.

# Armazenamento Flash - Tipos

- NOR Flash - Bits armazenados em portas NOR
  - Acesso randômico de leitura e escrita
  - Usado em memória para instruções em sistemas embarcados
- NAND flash - Bits armazenados em portas NAND
  - Mais denso, em termos de bits/área
  - Acesso por blocos, por exemplo, 2k, 4k, etc.
  - Acesso sequencial, similar aos HDDs
  - Mais baratos
  - Usados em USB keys, discos SSD, etc.

# Solid-State Disks (SSDs)

- "Discos" que usam memórias flash
- Interface compatível com as dos HDDs
- Páginas: 0,5 a 4 Kb
- Bloco: 32 a 128 págs.



Fonte: Mohamed Zahran (NYU) Computer Systems Organization slides

# SSD vs HDD

	Largura de Banda (Ac. Sequencial)	Custo/GB	Tamanho
HDD <sup>2</sup>	50-100 MB/s	\$0.03-0.07/GB	2-4 TB
SSD <sup>1,2</sup>	200-550 MB/s (SATA) 6 GB/s (Leitura PCI) 4.4 GB/s (Escrita PCI)	\$0.87-1.13/GB	200GB-1TB
DRAM <sup>2</sup>	10-16 GB/s	\$4-14*/GB	64GB-256GB

\*SK Hynix 9/4/13 fire

Fonte: <http://www.extremetech.com/computing/164677-storage-pricewatch-hard-drive-and-ssd-prices-drop-making-for-a-good-time-to-buy>

# SSD de 15 T !!!

- Samsung PM1633a 15.36TB SSD
- Inicio das vendas em Janeiro de 2016
- Interface SAS de 12Gb/s
- 2.5 polegadas



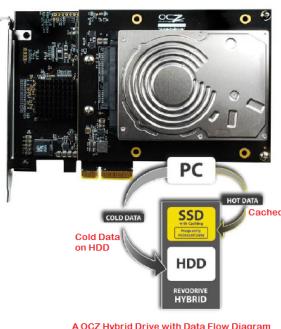
Samsung PM1633a MZILS15THMLS - solid state drive - 15.36 TB - SAS 12Gb/s - By NETCNA

by NETCNA

Be the first to review this item

Price: **\$15,299.95** & FREE Shipping

# Híbridos: SSD + HDD



# Você sabia??

- Fujio Masuoka inventou a memória flash em 1984, enquanto trabalhava para Toshiba. Capaz de ser apagada e reprogramada várias vezes, ganhou a indústria da memória do computador.
- Masuoka ficou descontente com a Toshiba por não reconhecer adequadamente seu trabalho e se demitiu para se tornar um professor da Universidade de Tohoku.
- Contrariando cultura de lealdade à empresa do Japão, processou seu ex-empregador recebendo em 2006 um pagamento único de ¥ 87m (\$ 758.000).



Fonte: <http://www.computerhistory.org/timeline/memory-storage/>

# BIG DATA

Vários dos slides são baseados nos slides originais do Professor Jimmy Lin da University of Waterloo (<https://cs.uwaterloo.ca/~jimmylin/>)

Big ??

- **Google™**
  - Processa 20 PB por dia (2008)
  - Coleta 20B páginas web pages por dia (2012)
  - Índice de busca tem 100+ PB (5/2014)
  - Bigtable serve 2+EB, 600M QPS (5/2014)

• **YAHOO!**

- Cluster de 330K nós, 365 PB (6/2014)

• **eBay™**

- Cluster de 10K nós, 150K núcleos, 150 PB (4/2014)

## Armazenamento X Processamento

- Conjuntos de dados de Terabytes são comuns e volumes na ordem de Petabytes começam a surgir com muita frequência.
- Tendência clara: Nossa capacidade de armazenar dados está rapidamente superando nossa habilidade de processar os dados que armazenamos.
- Mais preocupante: o aumento na capacidade de armazenamento está sobrepujando as melhorias em largura de banda. Está difícil até mesmo ler os dados que estão sendo armazenados.
  - Capacidade dos discos passou de dezenas de MB na medite dos 80 para alguns TB hoje em dia.
  - Por outro lado, a latência e a largura de banda melhoraram relativamente pouco
    - Latência melhorou 2X nos últimos 25 anos
    - Banda: talvez 50x

Big ??

• **facebook.**

- 300 PB data in Hive. +600 TB/day (4/2014)

•  **amazon web services**

- S3: 2T objects, 1.1M request/second

•  INTERNET ARCHIVE

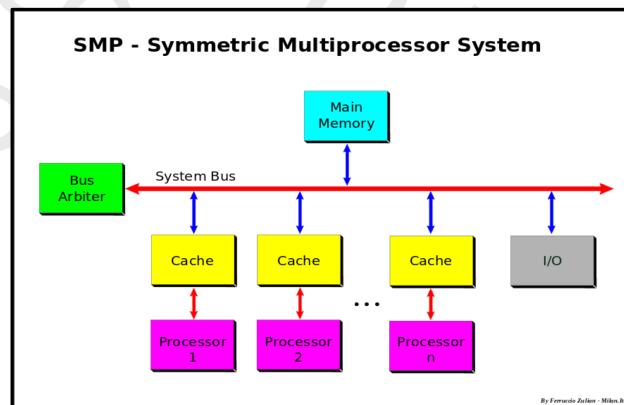
- 400B pages, 10+ PB (2/2014)

## Big???

- The Large Hadron Collider (CERN):
  - Maior acelerador de partículas e o de maior energia existente do mundo.
  - ~15 PB por ano
- Large Synoptic Survey Telescope (LSST) ~2020
  - Telescópio refletor com espelho primário de 8,4 metros, atualmente em construção. Vai fotografar o céu inteiro disponível em poucas algumas noites
  - 6-10 PB por ano
- Square Kilometre Array (SKA):
  - será o maior telescópio do mundo, capaz de captar ondas de rádio e que deve ficar pronto em 2017. Quando estiver terminado, vai estar espalhado pela Austrália, Nova Zelândia, África do Sul e outros países vizinhos como Moçambique e talvez Angola
  - 0.3 – 1.5 EB por ano (~2020)



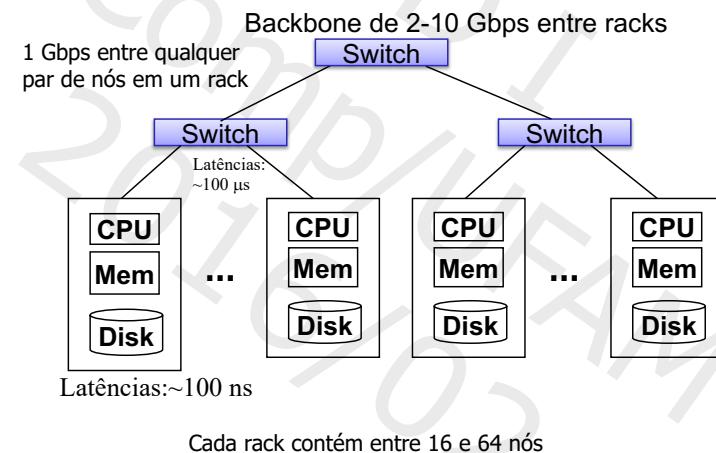
## Scaling “up” vs. “out”



## Scaling “up” vs. “out”

- Nenhuma máquina individual é “grande” o suficiente
  - Cluster Pequeno de Grandes Máquinas SMP
    - (16 máquinas de 128 núcleos)
      - X
  - Grande cluster de Máquinas Baratas
    - (128 máquinas de 16 núcleos)

## Scaling “up” vs. “out”





The datacenter *is* the computer!

Source: Google



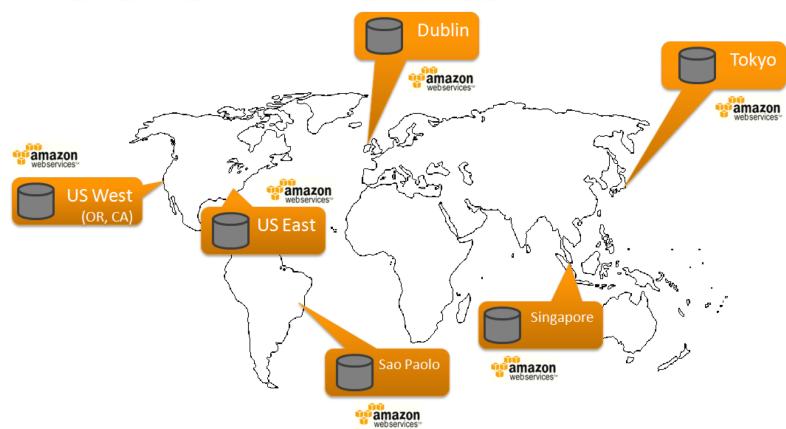
Source: Google



Source: Google

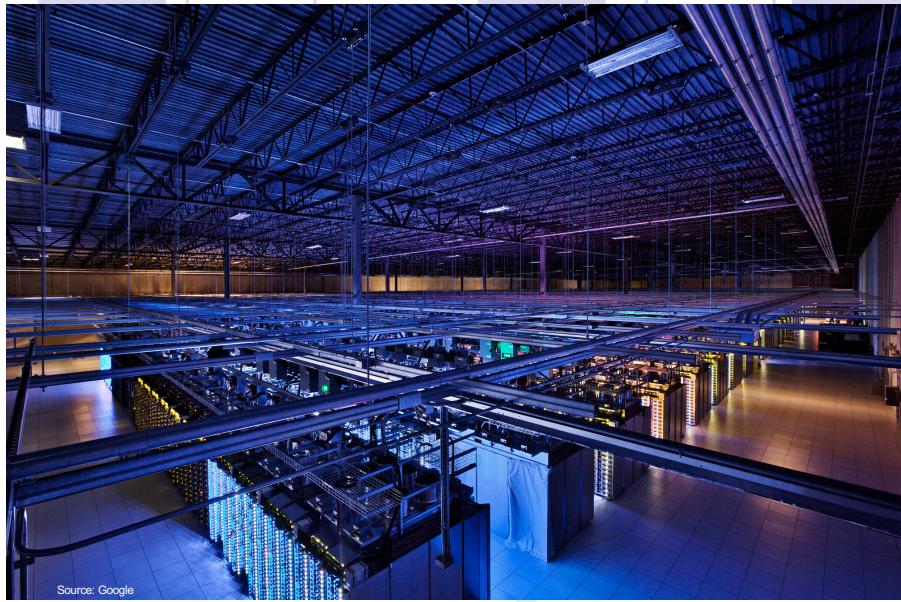


Source: Bonneville Power Administration



## Componentes Básicos





Source: Google

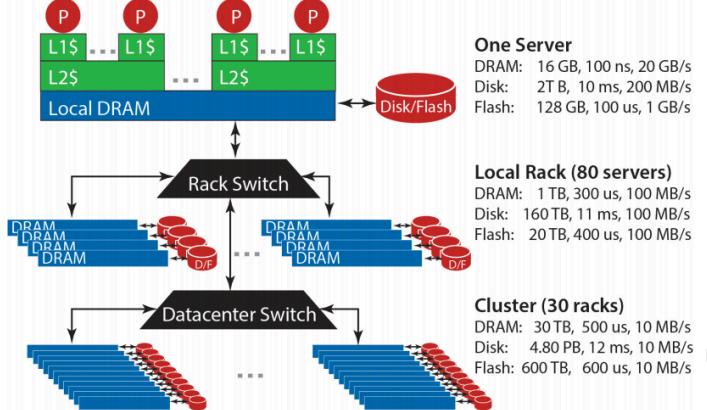


Source: Google



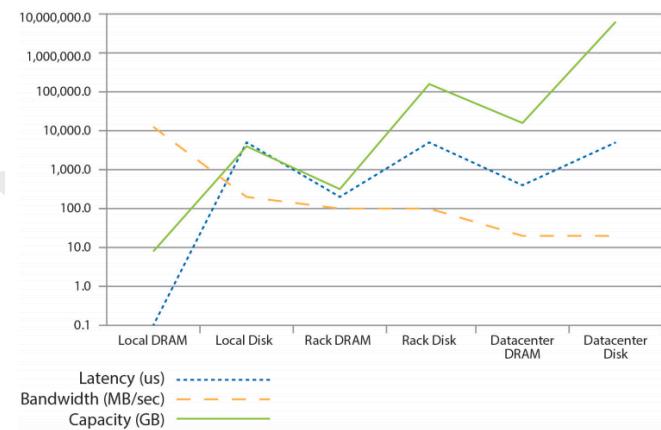
Source: Facebook

## Hierarquia de Armazenamento



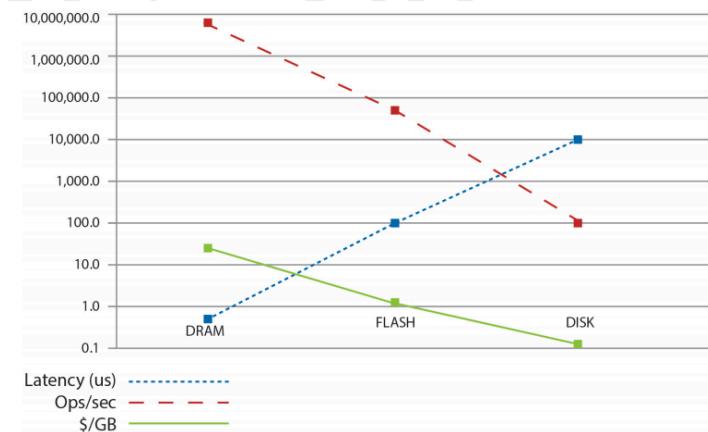
Fonte: Barroso and Urs Hözle (2013)

# Hierarquia de Armazenamento



Fonte: Barroso and Urs Hözle (2013)

# Hierarquia de Armazenamento



Fonte: Barroso and Urs Hözle (2013)