2018年度(平成30年度)版

Ver. 2018-10-11a

Course number: CSC.T363

コンピュータアーキテクチャ
Computer Architecture

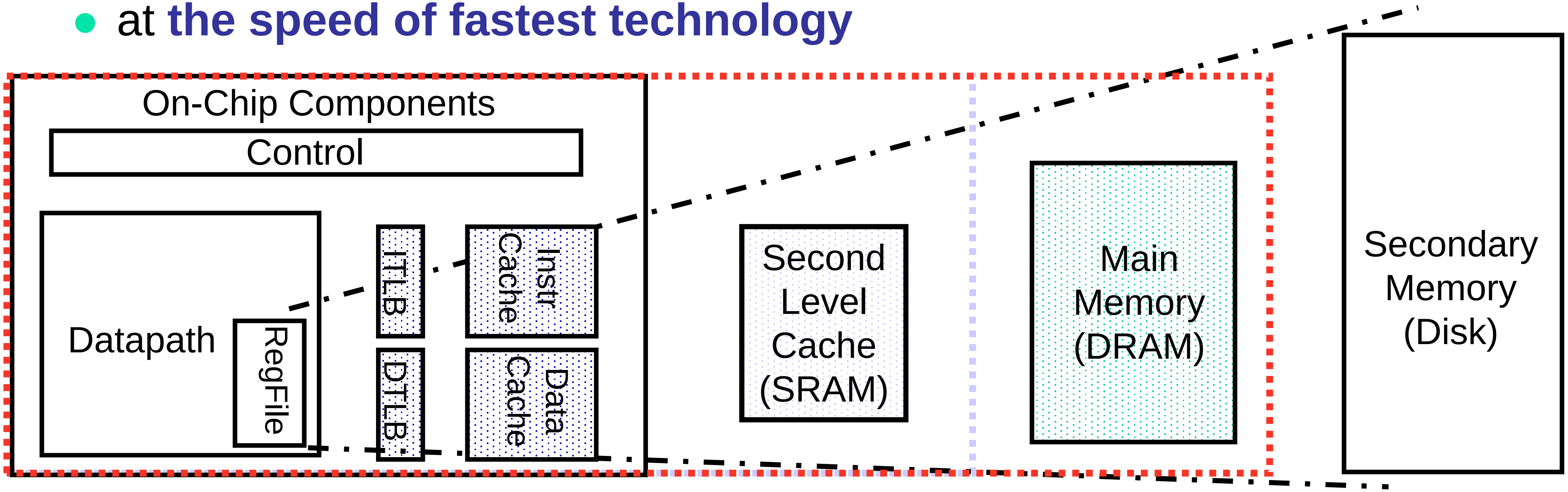
6. メモリシステムの階層化と信頼性
Memory Hierarchy and Dependability

www.arch.cs.titech.ac.jp/lecture/CA/
Room No.W321
Tue 13:20-16:20, Fri 13:20-14:50

吉瀬 謙二　情報工学系
Kenji Kise, Department of Computer Science
kise _at_ c.titech.ac.jp

A Typical Memory Hierarchy

- By taking advantage of **the principle of locality**（局所性）
  - Present **much memory** in **the cheapest technology**
  - at **the speed of fastest technology**

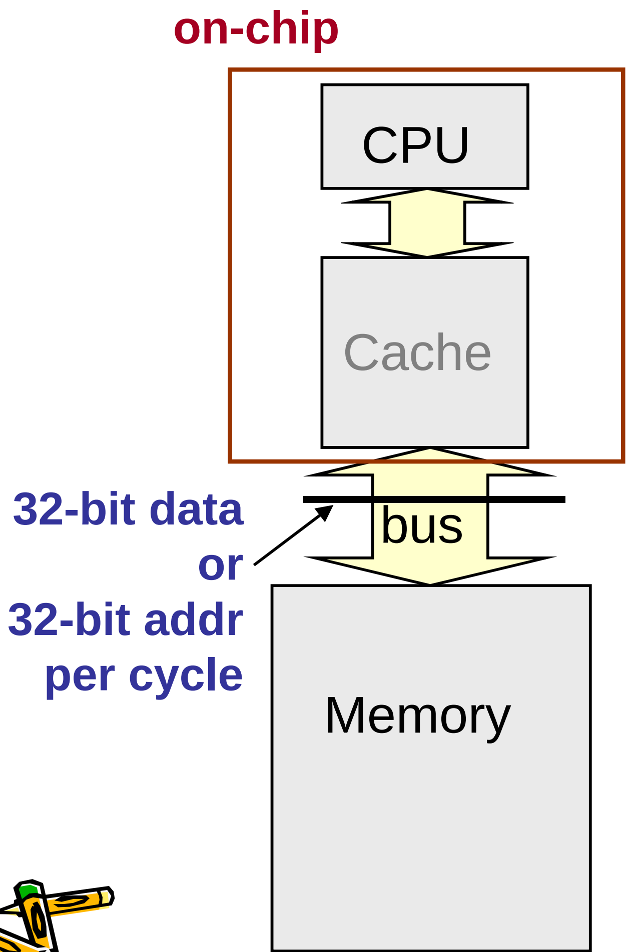


Speed (%cycles):	½'s	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	G's to T's
Cost:	highest				lowest

TLB: Translation Lookaside Buffer

Memory Systems that Support Caches

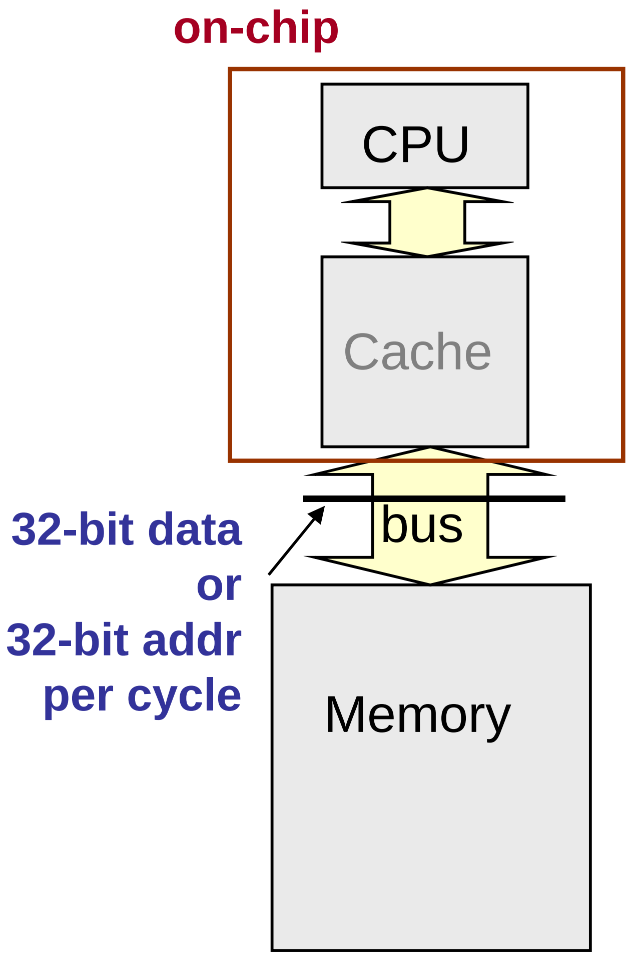
- The off-chip interconnect and memory architecture can affect overall system performance **in dramatic ways**



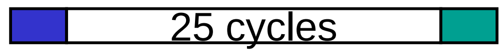
One word wide organization (one word wide bus and one word wide memory)

- **Assume**
  1. **1** clock cycle to send the address
  2. **25** clock cycles for DRAM **cycle time**, **8** clock cycles **access time**
  3. **1** clock cycle to return a word of data
- **Memory-Bus** to Cache **bandwidth**
  - **number of bytes transferred from memory to cache per clock cycle**

One Word Wide Memory Organization

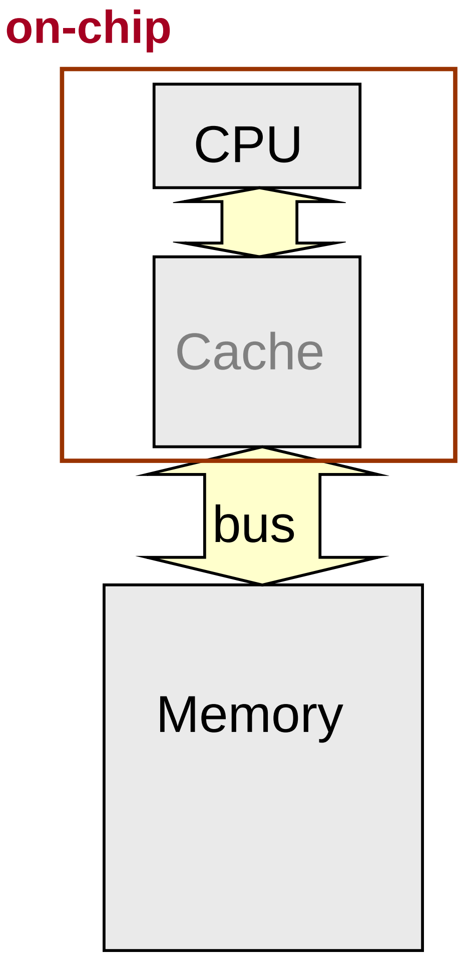


- The pipeline stalls the number of cycles for **one word** (32bit) from memory
  - 1 cycle to send address
  - 25 cycles to read DRAM
  - 1 cycle to return data
  - 27 **total clock cycles** miss penalty

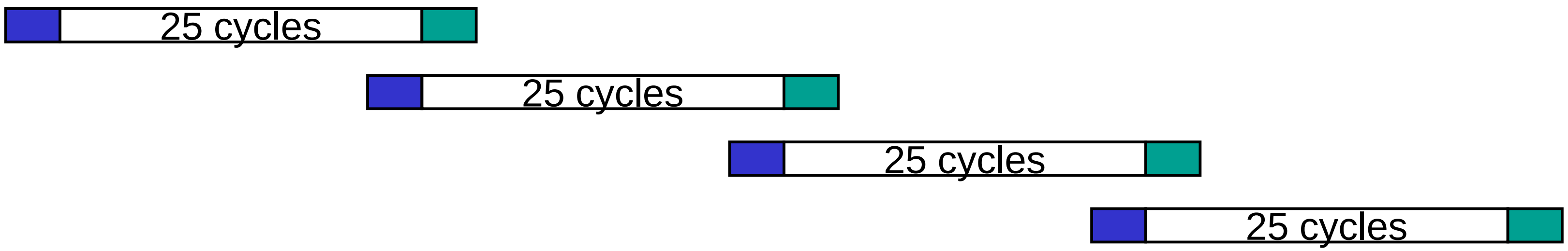


- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - 4 / 27 = 0.148 bytes per clock

One Word Wide Memory Organization, con't

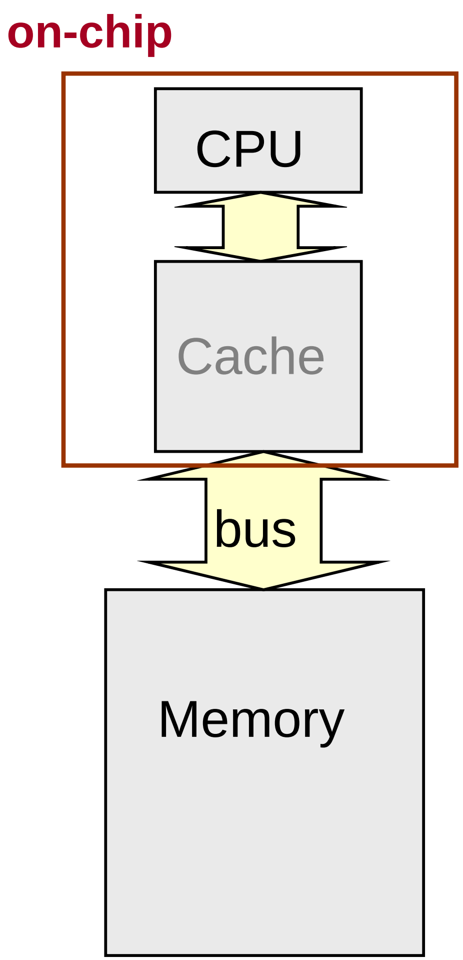


- What if the block size is **four words**?
  - 1 cycle to send 1st address
  - 4 * 25 = 100 cycles to read DRAM
  - 1 cycle to return last data word
  - 102 **total clock cycles** miss penalty

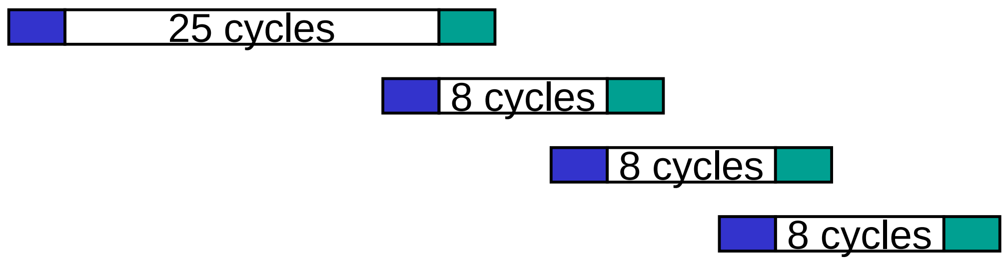


- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - (4 x 4) / 102 = 0.157 bytes per clock

One Word Wide Memory Organization, con't



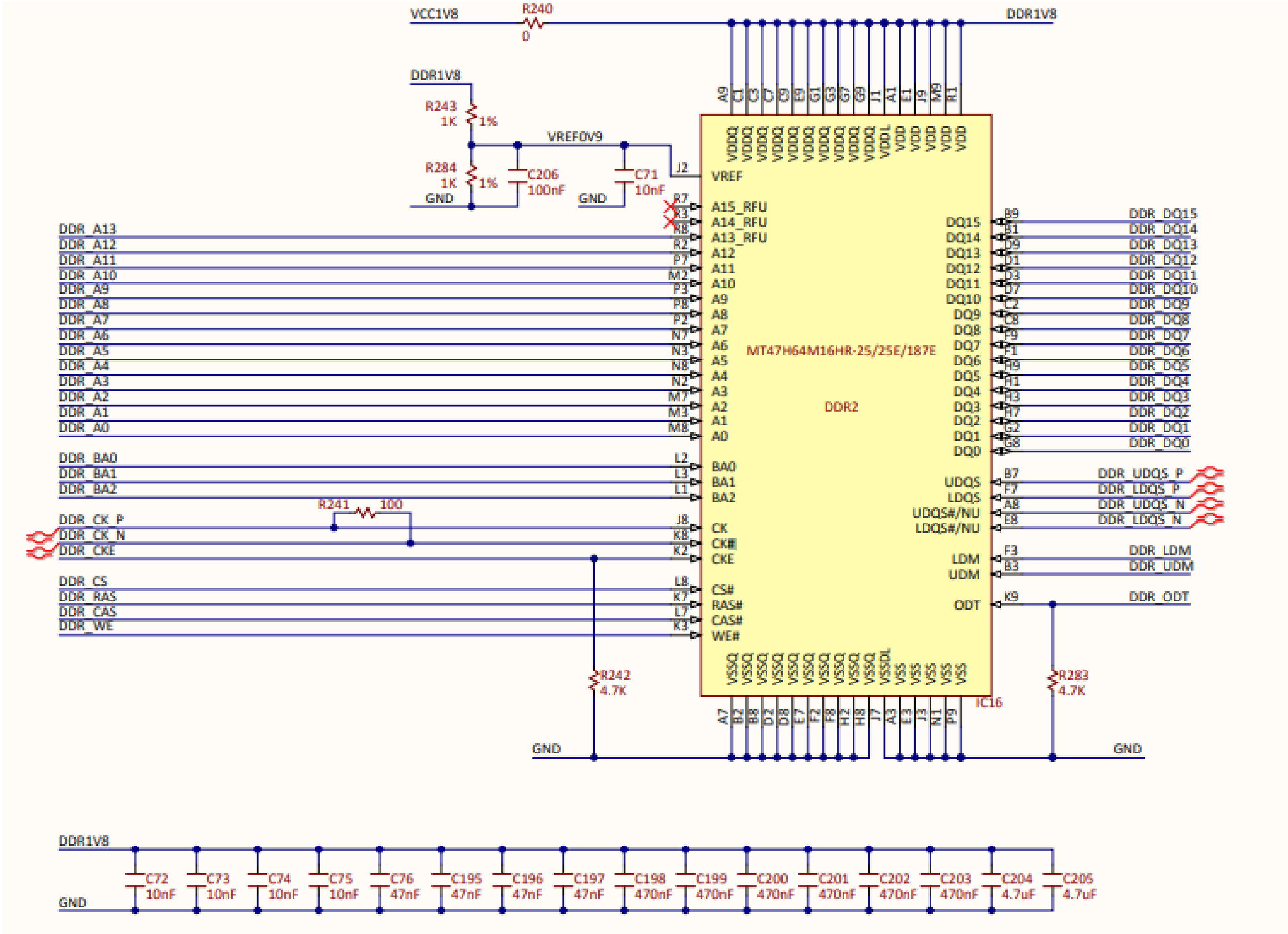
- What if the block size is **four words** and if a **page mode DRAM** is used?
  - 1 cycle to send 1st address
  - 25 + (3 * 8) = 49 cycles to read DRAM
  - 1 cycle to return last data word
  - 51 **total clock cycles** miss penalty



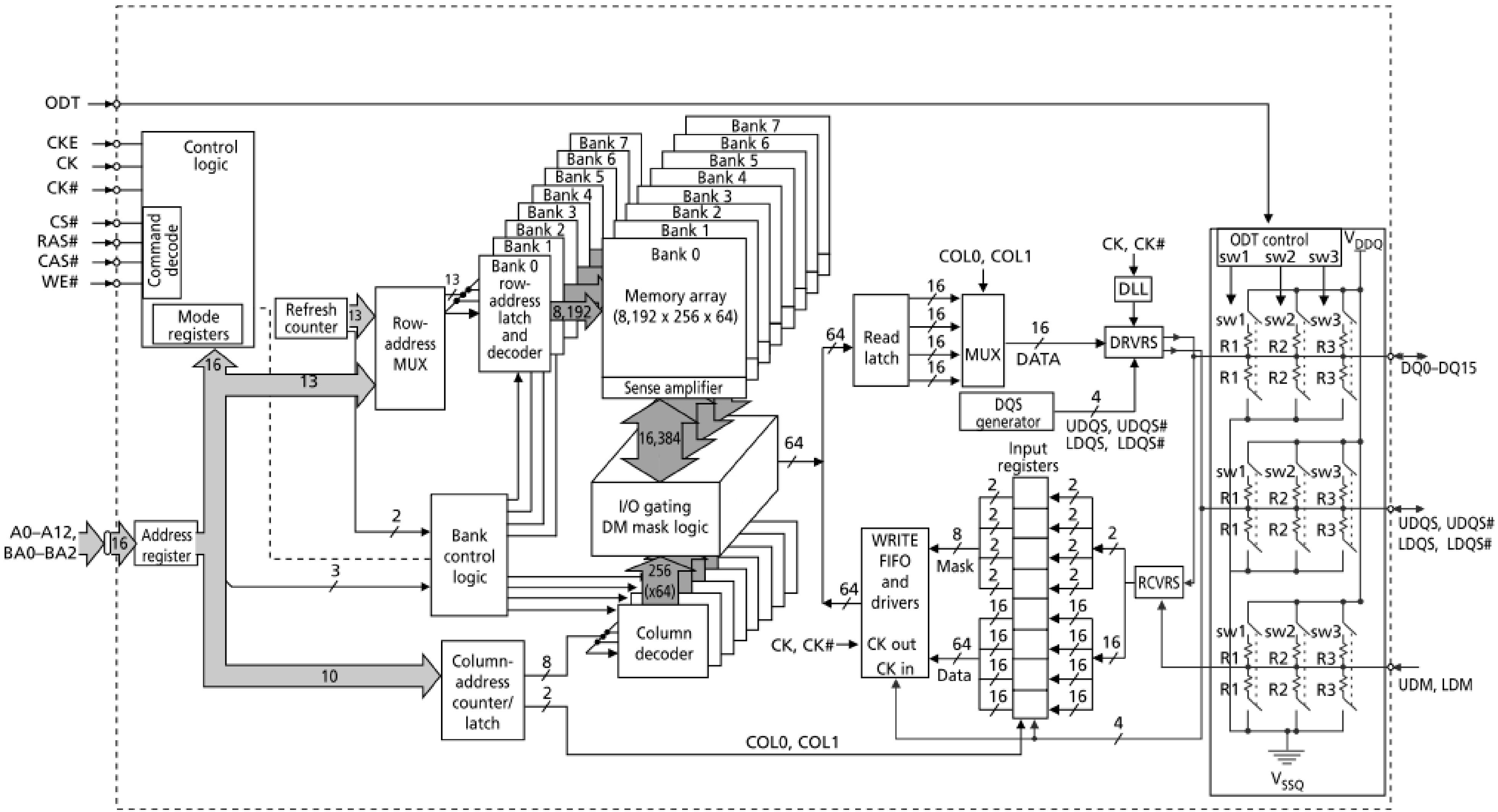
- Number of bytes transferred per clock cycle (**bandwidth**) for a single miss
  - (4 x 4) / 51 = 0.314 bytes per clock

NEXYS 4 DDR

- Micron MT47H64M16HR-25:H DDR2 memory
  - 128MiB DDR2, 16-bit wide interface

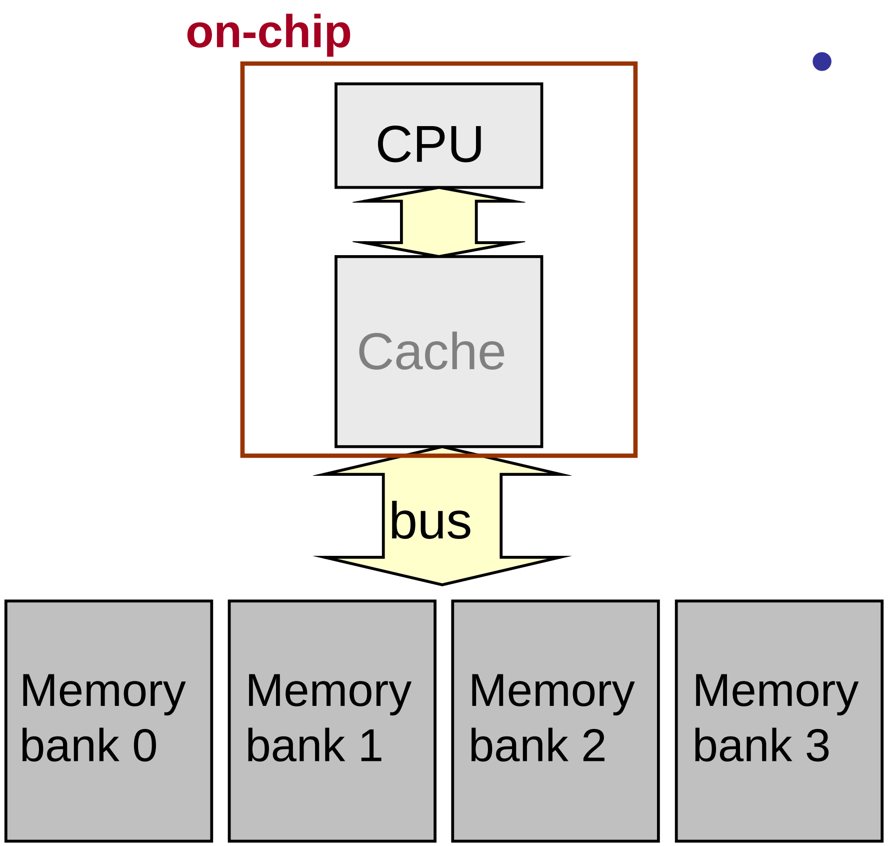


Micron MT47H64M16HR-25:H



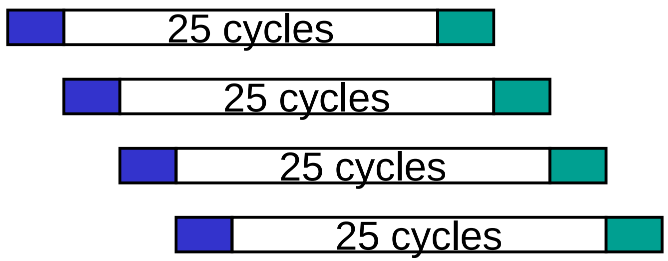
Micron datasheet

Interleaved(インターリーブ) Memory Organization



With **parallelism**

- For a block size of **four words with interleaved memory (4 banks)**
  - **1** cycle to send 1st address
  - **25 + 3 = 28** cycles to read DRAM
  - **1** cycle to return last data word
  - **30 total clock cycles** miss penalty



- Number of bytes transferred per clock cycle (bandwidth) for a single miss
  - **(4 x 4) / 30 = 0.533 bytes per clock**

The Memory System's **Fact and Goal**

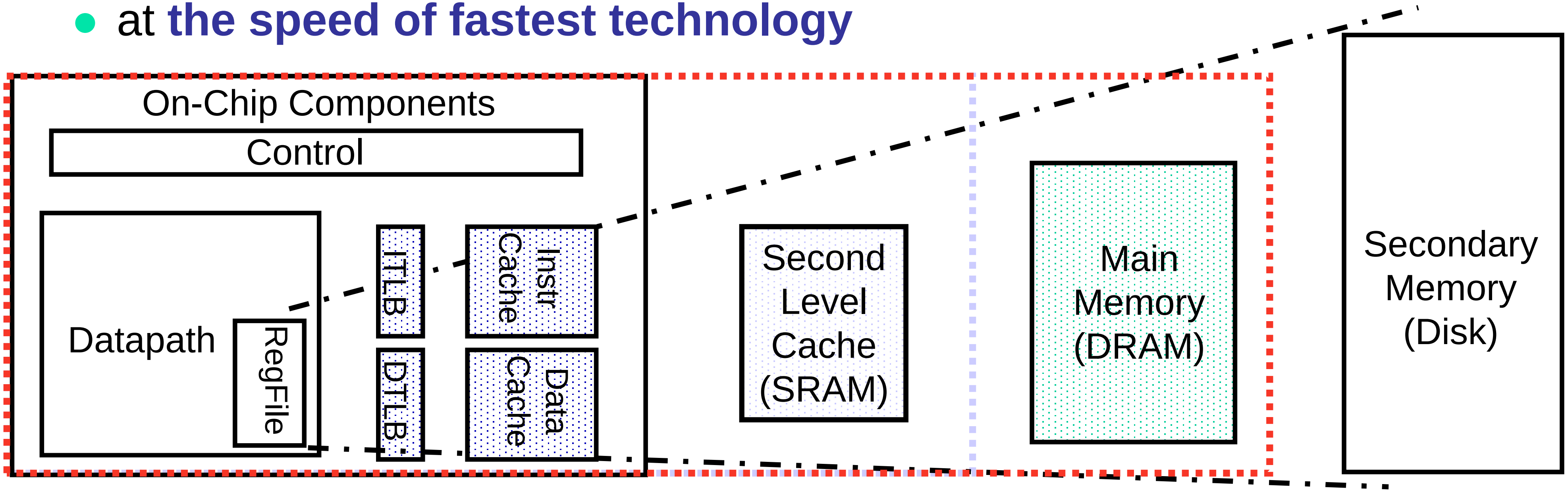
- **Fact:**
  Large memories are slow and
  fast memories are small

- How do we create a memory that gives the **illusion** of being large, cheap and fast ?

  - With **hierarchy** (階層)
  - With **parallelism** (並列性)

A Typical Memory Hierarchy

- By taking advantage of **the principle of locality**（局所性）
  - Present **much memory** in **the cheapest technology**
  - at **the speed of fastest technology**

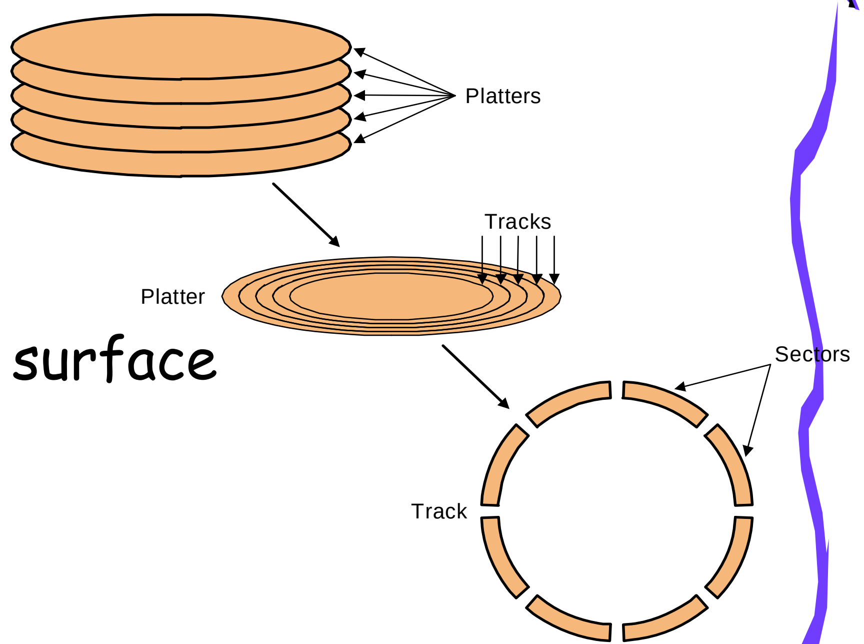


Speed (%cycles):	½'s	1's	10's	100's	1,000's
Size (bytes):	100's	K's	10K's	M's	G's to T's
Cost:	highest				lowest

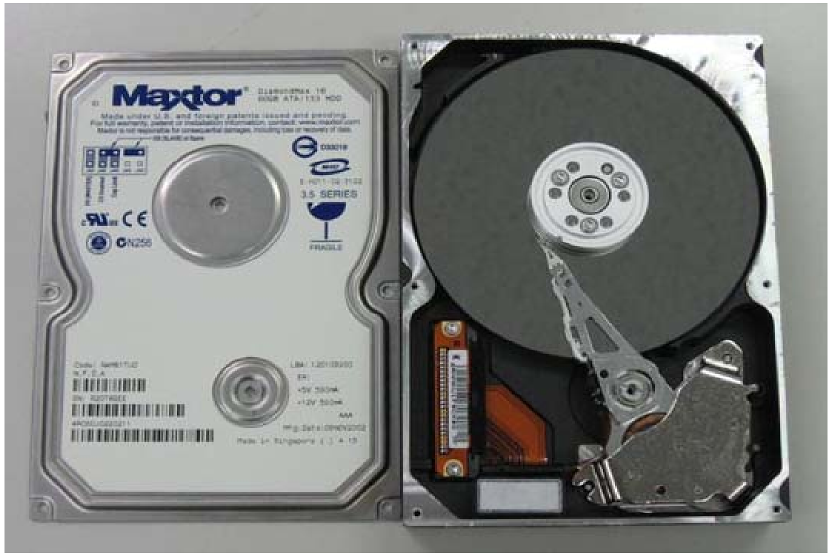
TLB: Translation Lookaside Buffer

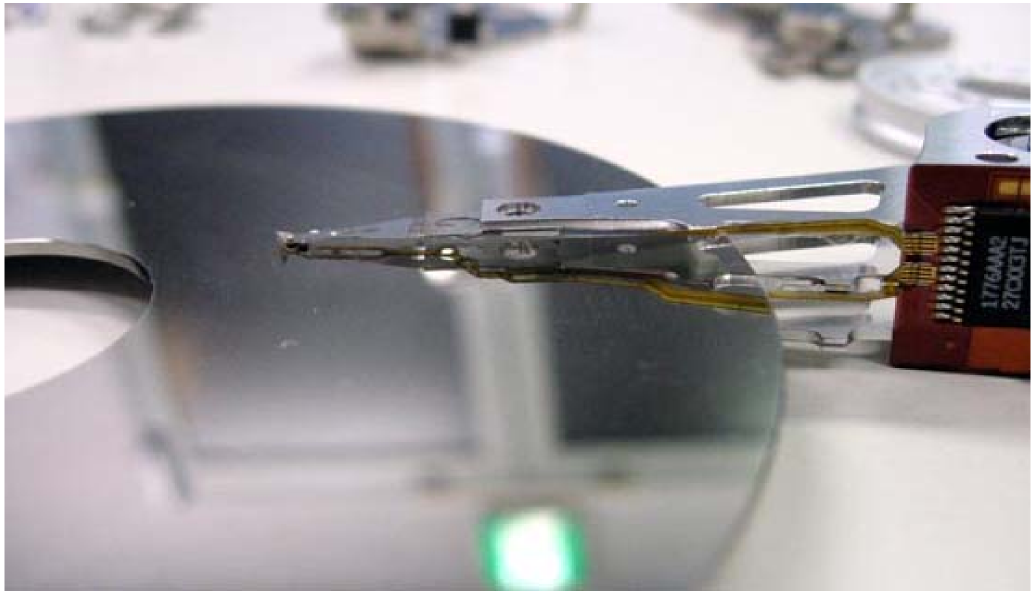
Magnetic Disk（磁気ディスク）

- Purpose
  - Long term, **nonvolatile**（不揮発性）storage
  - Lowest level in the memory hierarchy
    - slow, large, inexpensive
- General structure
  - A rotating **platter** coated with a magnetic surface
  - A moveable read/write **head** to access the information on the disk
- Typical numbers
  - 1 to 4 platters per disk of 1" to 5.25" in diameter (3.5" dominate in 2004)
  - Rotational speeds of 5,400 to 15,000 RPM (rotation per minute)
  - 10,000 to 50,000 **tracks** per surface
    - **cylinder** - all the tracks under the head at a given point on all surfaces
  - 100 to 500 **sectors** per track
    - the smallest unit that can be read/written (typically **512B**)



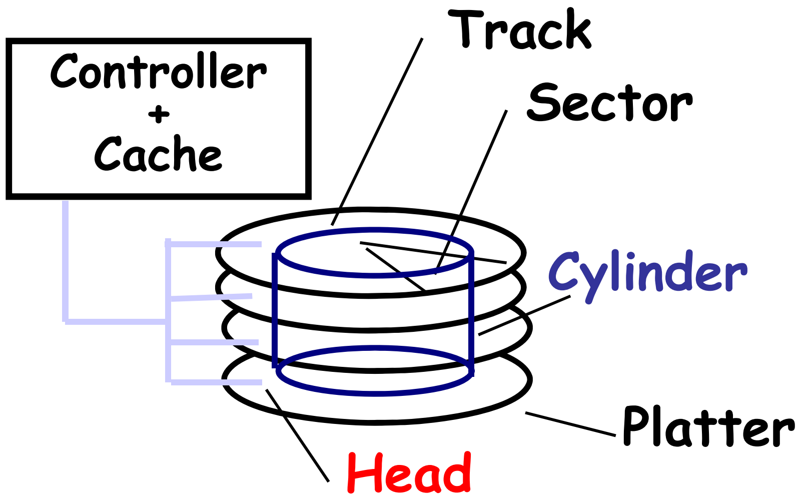
Magnetic Disk（磁気ディスク）

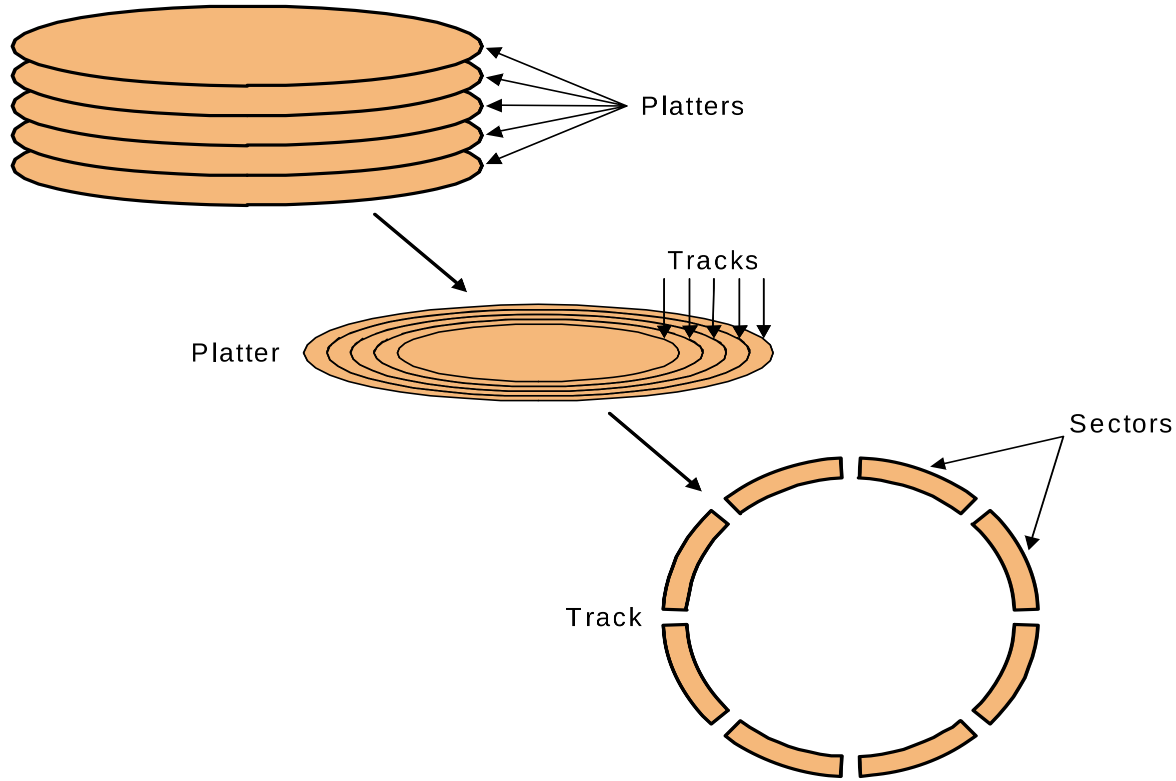




http://sougo057.aicomp.jp/0001.html

Disk Drives



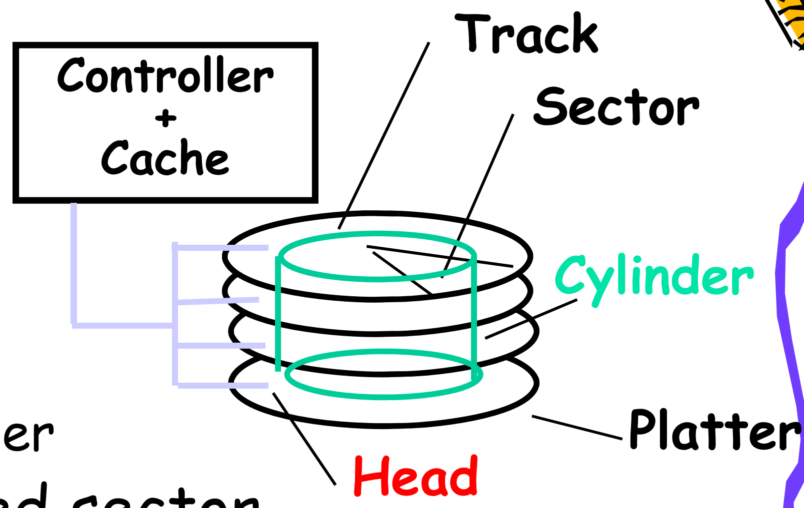


To access data:

- seek time (シーク時間): position the head over the proper track
- rotational latency (回転待ち時間): wait for desired sector
- transfer time (転送時間): grab the data (one or more sectors)
- Controller time(制御時間): the overhead the disk controller imposes in performing a disk I/O access

Magnetic Disk Characteristic

- **Disk read/write components**
  1. **Seek time**: position the head over the proper track (**3 to 14 ms avg**)
     - due to locality of disk references the actual average seek time may be only 25% to 33% of the advertised number
  2. **Rotational latency**: wait for the desired sector to rotate under the head (½ of 1/RPM converted to ms)
     - 0.5/5400RPM = 0.5/90 rotations per second = **5.6 ms**
     - 0.5/15000RPM = 0.5/250 rotations per second = **2.0 ms**
  3. **Transfer time**: transfer a block of bits (one or more sectors) under the head to the disk controller's cache (**30 to 80 MB/s** are typical disk transfer rates)
  4. **Controller time**: the overhead the disk controller imposes in performing a disk I/O access (**typically < .2 ms**)



Typical Disk Access Time

- The average time to read or write a 512B sector for a disk rotating at 10,000RPM with average seek time of 6ms, a 50MB/sec transfer rate, and a 0.2ms controller overhead

Avg disk read/write time

= 6.0ms + 0.5/(10000RPM/(60sec/minute))+ 0.5KB/(50MB/sec) + 0.2ms

= 6.0 + 3.0 + 0.01 + 0.2

= 9.21ms

If the measured average seek time is **25%** of the advertised average seek time, then

Avg disk read/write = 1.5 + 3.0 + 0.01 + 0.2 = 4.71ms

- The **rotational latency** is usually the largest component of the access time

Disk Latency & Bandwidth Milestones

- Disk **latency** is one average seek time plus the rotational latency.
- Disk **bandwidth** is the peak transfer time of formatted data from the media (not from the cache).

	CDC Wren	SG ST41	SG ST15	SG ST39	SG ST37
Speed (RPM)	3600	5400	7200	10000	15000
Year	**1983**	**1990**	**1994**	**1998**	**2003**
Capacity (Gbytes)	0.03	1.4	4.3	9.1	73.4
Diameter (inches)	5.25	5.25	3.5	3.0	2.5
Interface	ST-412	SCSI	SCSI	SCSI	SCSI
Bandwidth (MB/s)	0.6	4	9	24	86
Latency (msec)	48.3	17.1	12.7	8.8	5.7

Patterson, CACM Vol 47, #10, 2004

Reliability（信頼性）, Availability

- Reliability – measured by the mean time to failure (平均故障時間, MTTF).
- Service interruption is measured by mean time to repair (平均修復時間, MTTR)
- Availability（アベイラビリティ）

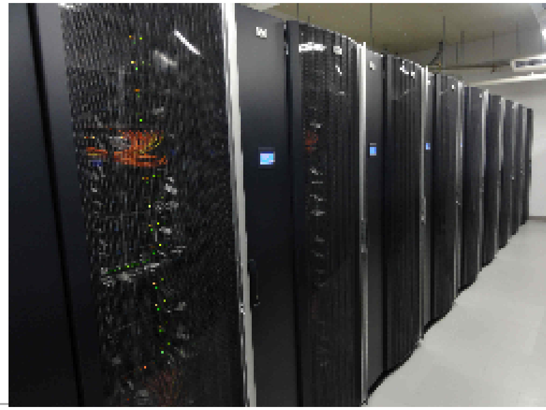
$$\text{Availability} = \text{MTTF} / (\text{MTTF} + \text{MTTR})$$

- To increase MTTF, either improve the quality of the components or <u>design the system to continue operating in the presence of faulty components</u>
  1. Fault avoidance: preventing fault occurrence by construction
  2. Fault tolerance: using redundancy to correct or bypass faulty components (hardware)

高信頼ディスクの典型的なMTTF は100万時間 (114年) 程度

TSUBAME 2.0

TSUBAME2では上記のとおり3種類の計算ノードを提供していますが、そのほとんどは54GBのメモリを搭載したThinノードです。共有メモリとして54GB以上用いる特定の場合をのぞいて最新GPUを搭載したThinノードをお使いください。Thinノードの計算性能はCPUが2基合計で153GFlops(ターボブースト時)、GPUが3基合計で1545GFlopsです(CPU、GPU共に倍精度浮動小数点演算性能)。またメモリバンド幅はCPU側が2CPU合算で64GB/s、GPU側が3基合算で462GB/sになります。それぞれハードウェアが出しうる理論ピーク性能であり実際のアプリケーションでの性能はこれに劣りますが、TSUBAME2ではCPU性能に比べてGPU性能を重視した構成になっております。



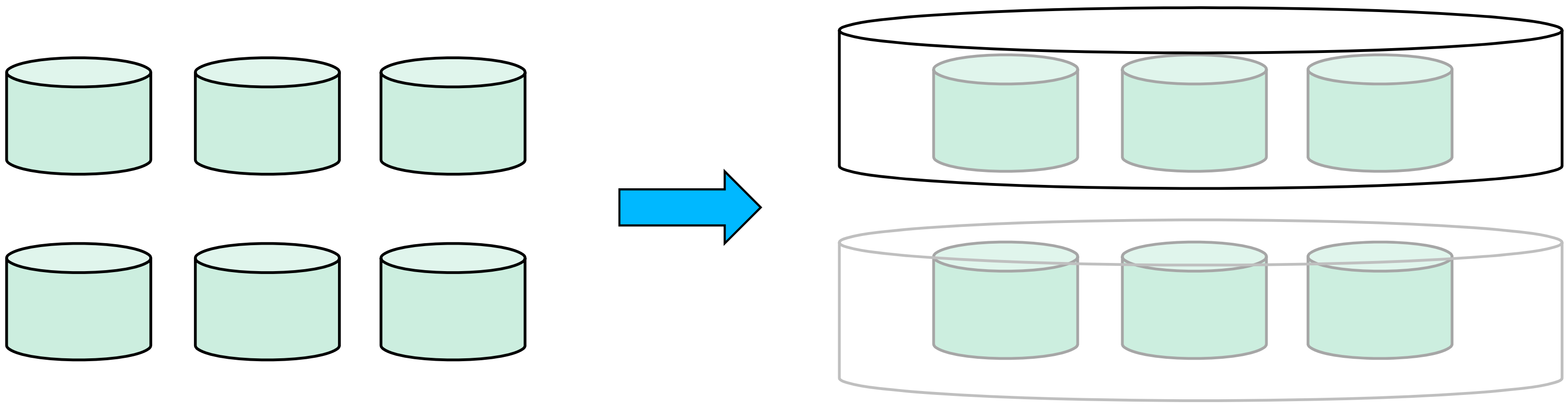
ストレージ

TSUBAME2ではストレージ領域としてホームディレクトリを提供するホーム領域と大規模データ処理用並列ファイルシステム領域の2種類のストレージ領域が利用可能であり、さらにテープドライブによる遠隔バックアップによる障害対策がとられています。

ストレージ種別	用途	プロトコル	構成	マウント先
ホーム	計算ノードのホームディレクトリ用 (NFS)、学内ストレージサービス (CIFS)、学内ホスティングサービス (iSCSI)	NFS, CIFS, iSCSI	BlueArc Mercury 100 (一部GRIDScalar)	/home
並列ファイルシステム領域	大規模データ処理用、実行時の中間データなどのためのスクラッチ領域	Lustre GPFS	MDS: HP DL360 G6 x 6, OSS: HP DL360 G6 x 20, DDN SFA 10K x 3, 2TB SATA x 3550, 600GB SAS x 50	/work0 /gscr0 /data0

RAID: Redundant Array of Inexpensive Disks

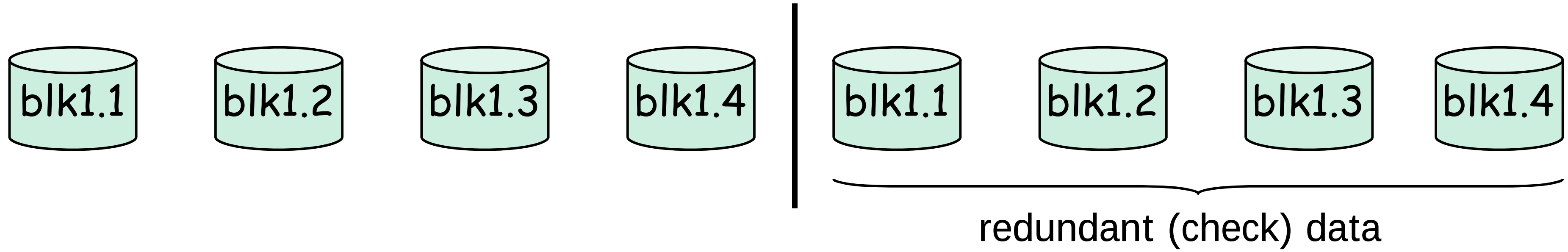
- Arrays of small and inexpensive disks
  - Increase potential throughput by having many disk drives
  - Data is spread over multiple disk
  - Multiple accesses are made to several disks at a time
- **Reliability** is lower than a single disk
- But **availability** can be improved by adding redundant disks



RAID: Level 0 (RAID 0, 冗長性なし, ストライピング)

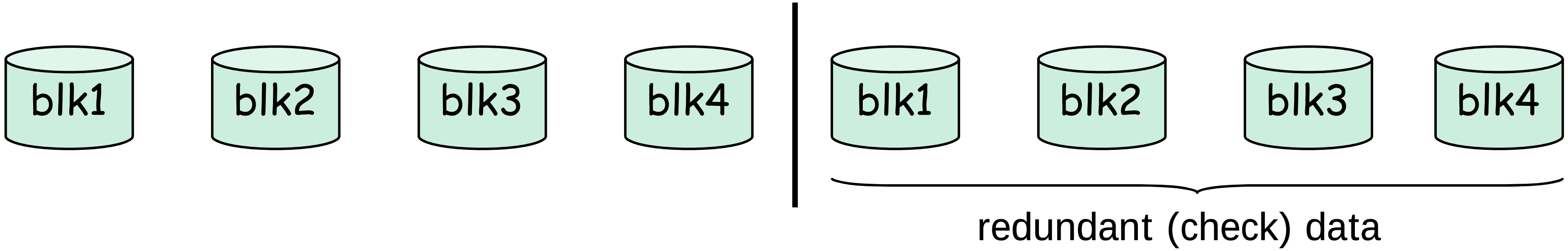
- Multiple smaller disks as opposed to **one big disk**
  - Spreading the blocks over multiple disks – **striping** – means that multiple blocks can be accessed in parallel increasing the performance
  - 4 disk system gives four times the throughput of a 1 disk system
  - Same cost as one *big* disk – assuming 4 small disks cost the same as one big disk
- No redundancy, so what if one disk fails?

RAID: Level 1 (Redundancy via Mirroring)



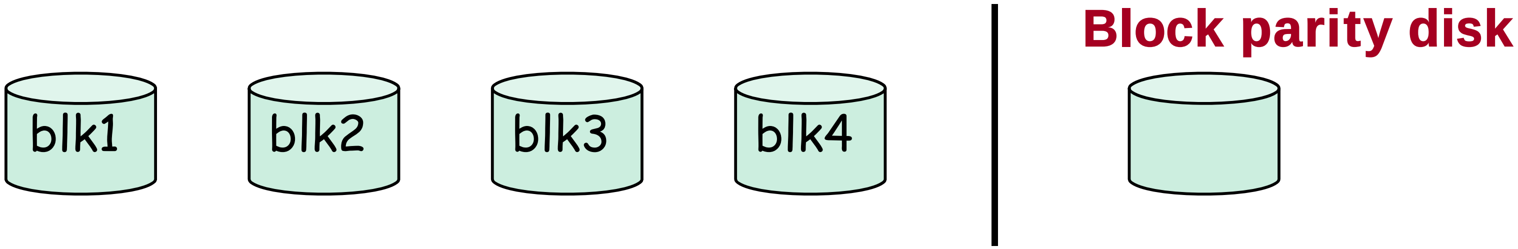
- Uses twice as many disks for redundancy so there are always two copies of the data
  - The number of redundant disks = the number of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only 1/2 the performance of RAID 0
- What if one disk fails?
  - If a disk fails, the system just goes to the "**mirror**" for the data

RAID: Level 0+1 (RAID01, Striping with Mirroring)



- Combines the best of RAID 0 and RAID 1, data is striped across four disks and mirrored to four disks
  - Four times the throughput (due to striping)
  - # redundant disks = # of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only 1/2 the performance of RAID 0
- What if one disk fails?
  - If a disk fails, the system just goes to the “**mirror**” for the data

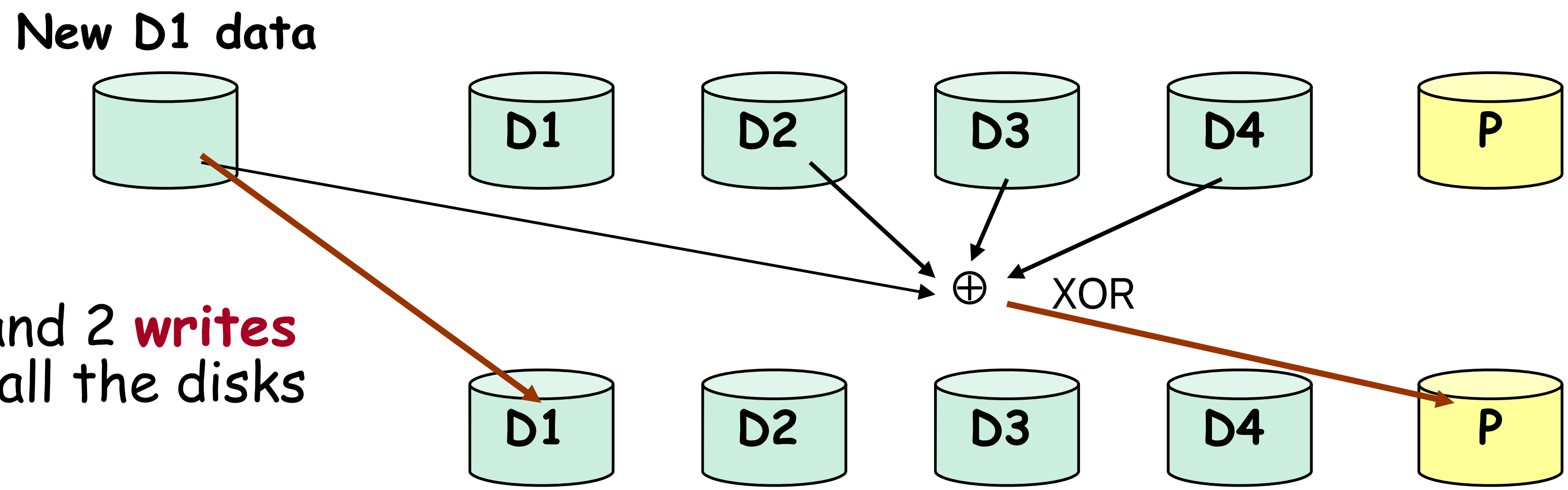
RAID: Level 4 (Block-Interleaved Parity)



- Cost of higher availability still only 1/N but the parity is stored as **blocks** associated with sets of data blocks
  - Four times the throughput (striping)
  - # redundant disks = 1 × # of protection groups
  - Supports "small reads" and "small writes"
    (reads and writes that go to just one (or a few) data disk in a protection group)

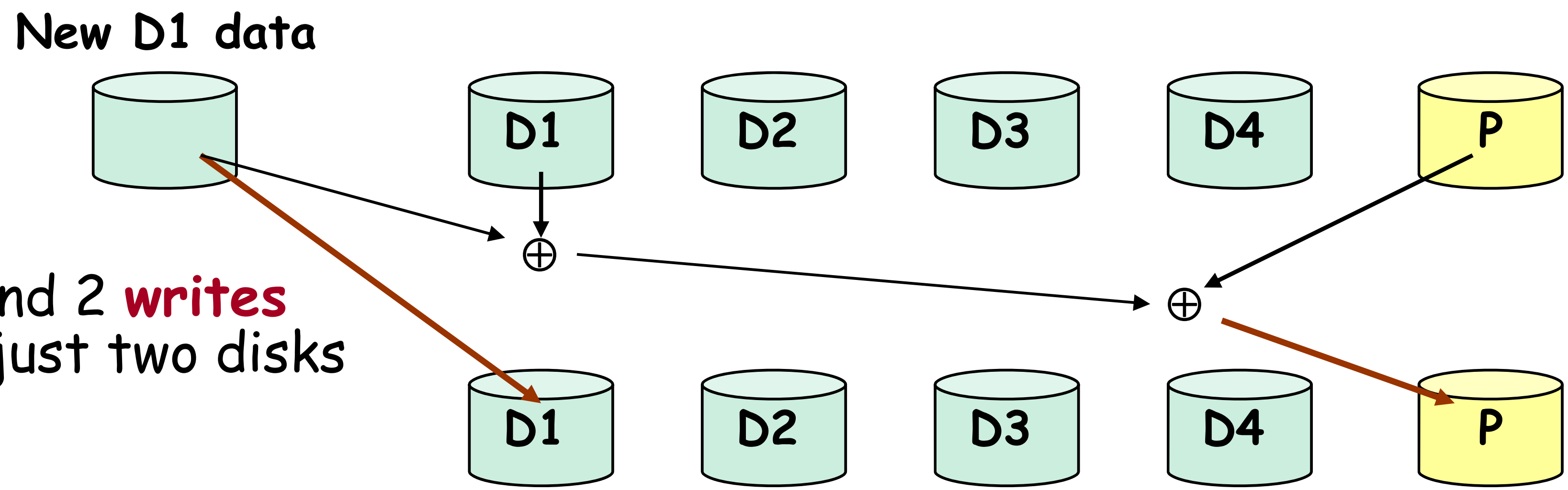
Small Reads and Small Writes

- RAID 3



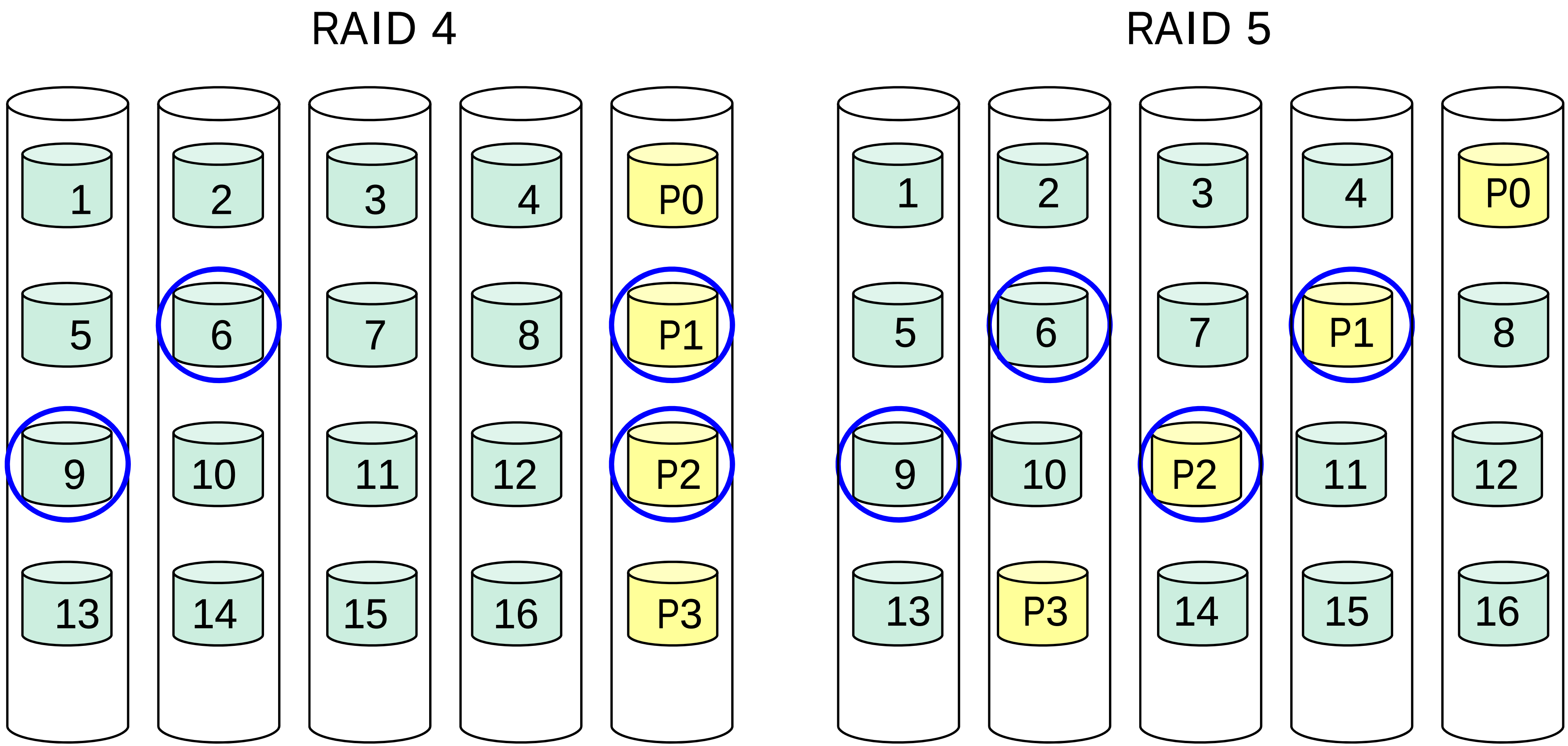
3 **reads** and 2 **writes** involving all the disks

- RAID 4 small reads and small writes



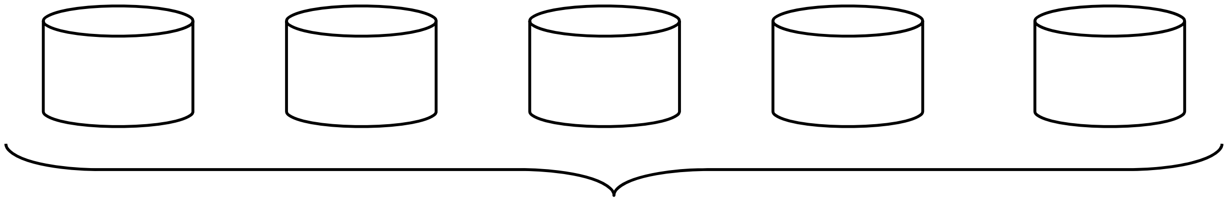
2 **reads** and 2 **writes** involving just two disks

Distributing Parity Blocks



- By distributing parity blocks to all disks, some small writes can be performed **in parallel**

RAID: Level 5 (Distributed Block-Interleaved Parity)



one of these assigned as the block parity disk

- Cost of higher availability still only 1/N but the parity block can be located on any of the disks
  so there is no single bottleneck for writes
  - Still four times the throughput (striping)
  - # redundant disks = 1 × # of protection groups
  - Supports "**small reads**" and "**small writes**" (reads and writes that go to just one (or a few) data disk in a protection group)
  - Allows **multiple simultaneous writes**