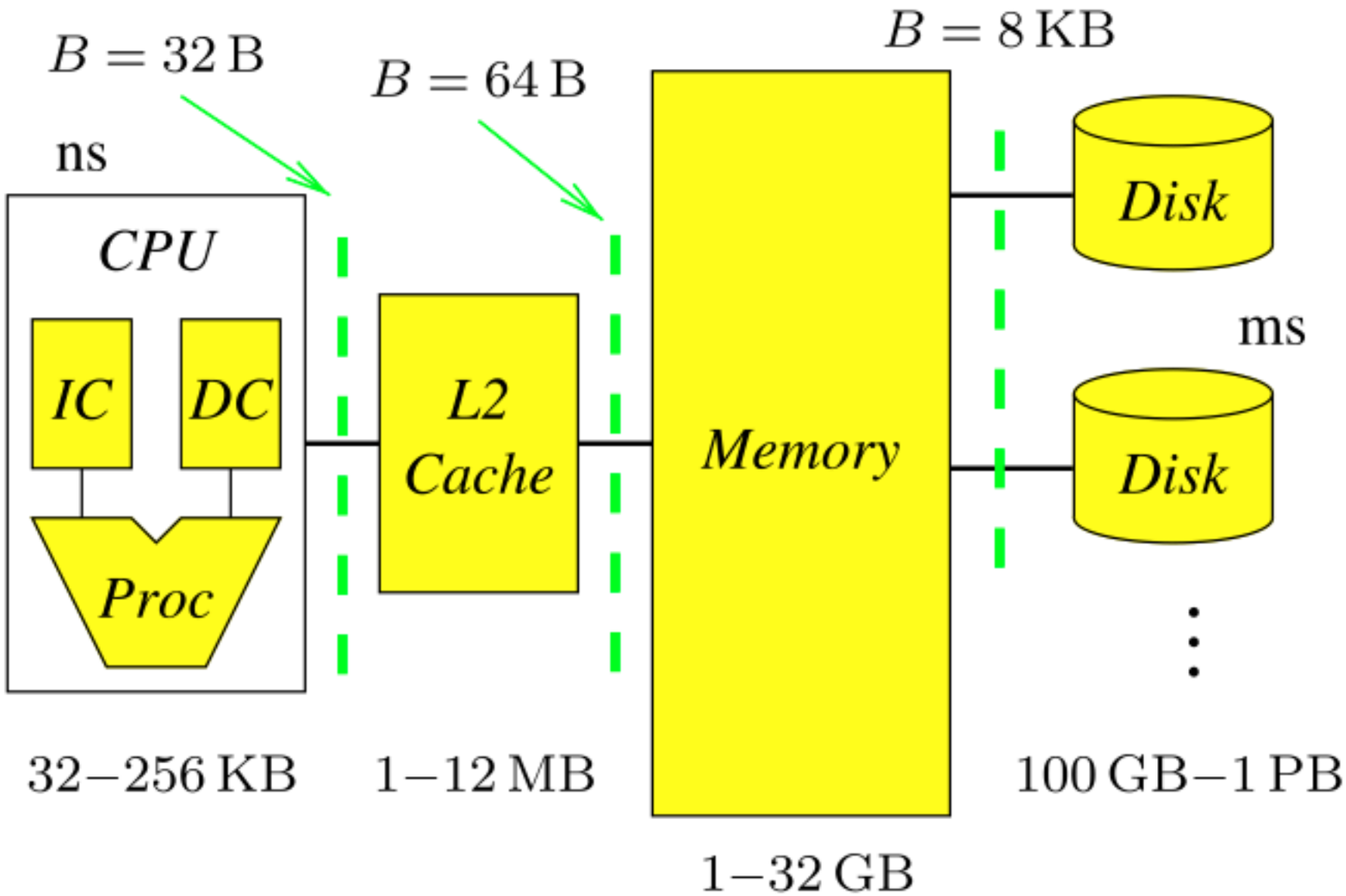
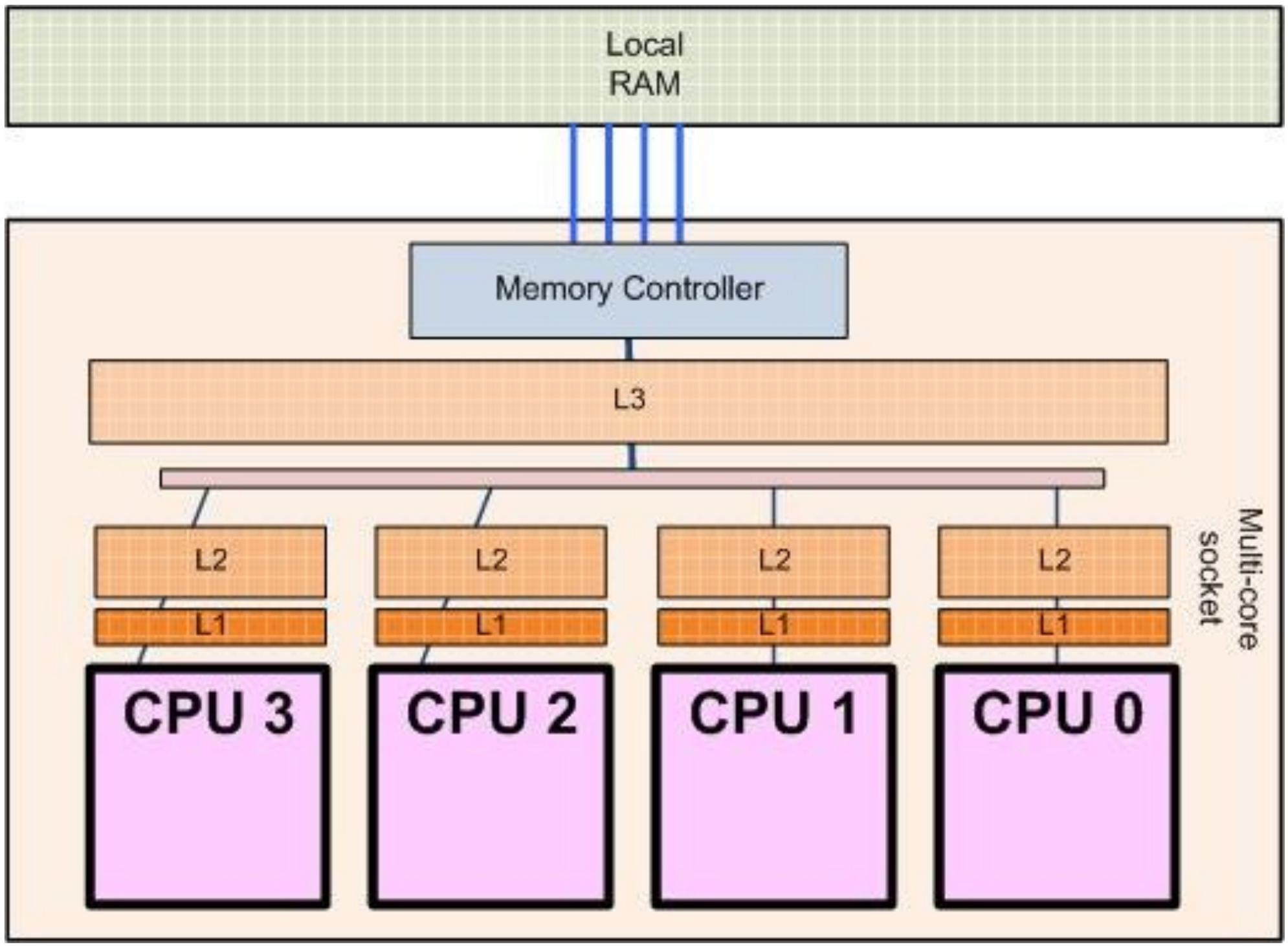
External Memory Algorithms

The Memory Hierarchy

B=Block size

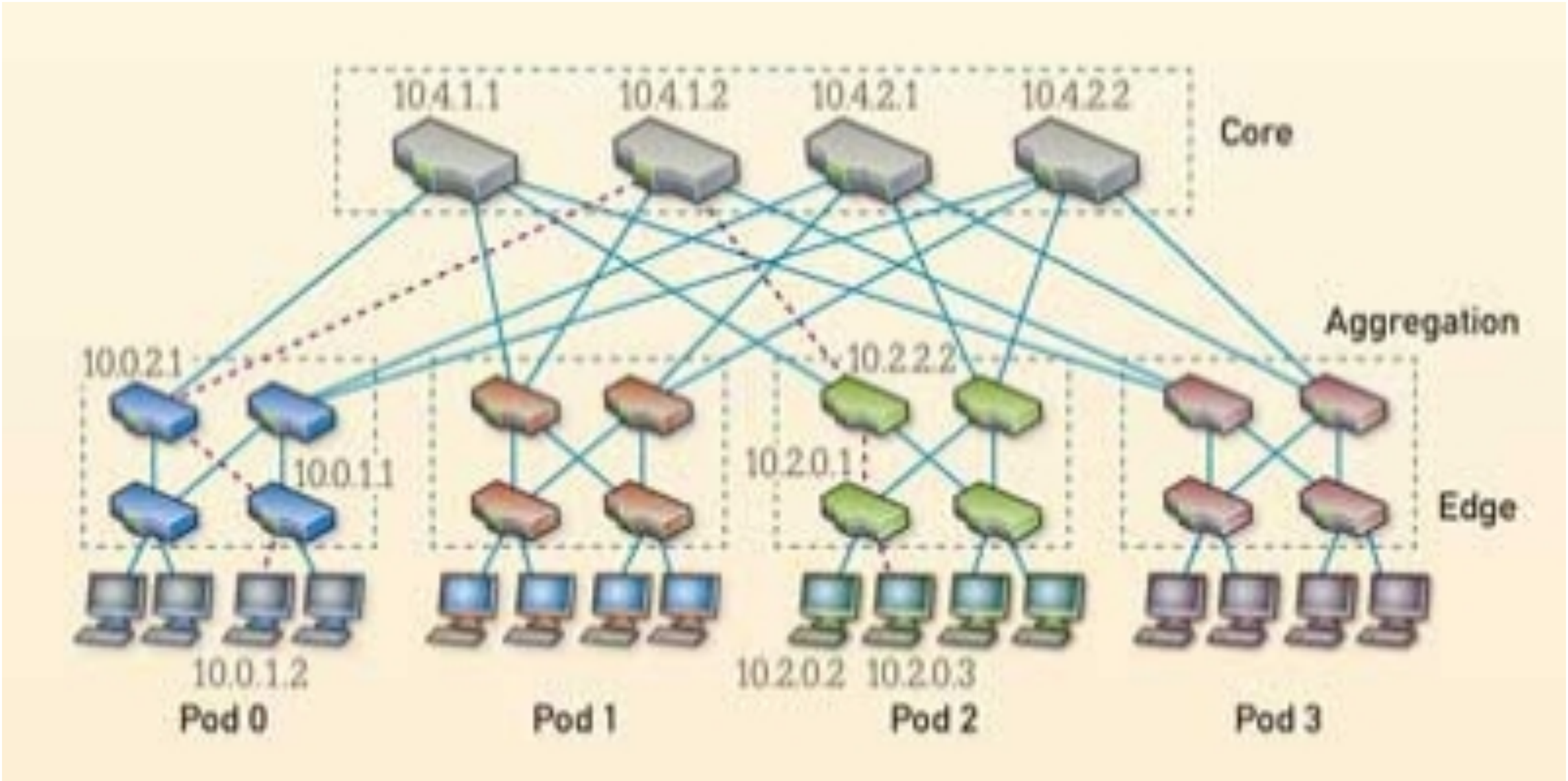


Multi-core computers



Data-Center networks

Taken from the academic paper authored by UCSD's Mohammad Al-Fares, Alexander Loukissas and Amin Vahdat, this illustration of a simple fat-tree architecture shows the path (dotted line) that packets would follow from one server to another. Source: A Scalable, Commodity Data Center Network Architecture.



Data-centers

Up to date Characteristics

	Cost	Power	Block Size	Bandwidth
L1-L2	On-Chip with CPU	low-end: iTouch: 1Watt high-end: Intel Core i7-950: 130Watt	L1:32-64 Bytes L2: 64-256 Bytes	100s of GB/ sec
DRAM	8-16/GB	1-2 Watts/GB	max throughput at: 8-16KB	50-70 GB/sec
SSD	High end, $11/GB. Low end: $1-4/GB About $50,000 for ION	high end: 0.15W/GB Low End: 0.05 W/GB	4KB	high end: 1-3GB/sec low end: .1-.2GB/sec
Disk	0.03-0.1$/GB	0.01W/GB	4KB	100MB/sec sequential 0.4-2.0MB/sec random Getting to each block takes 2-10ms

The disk drive

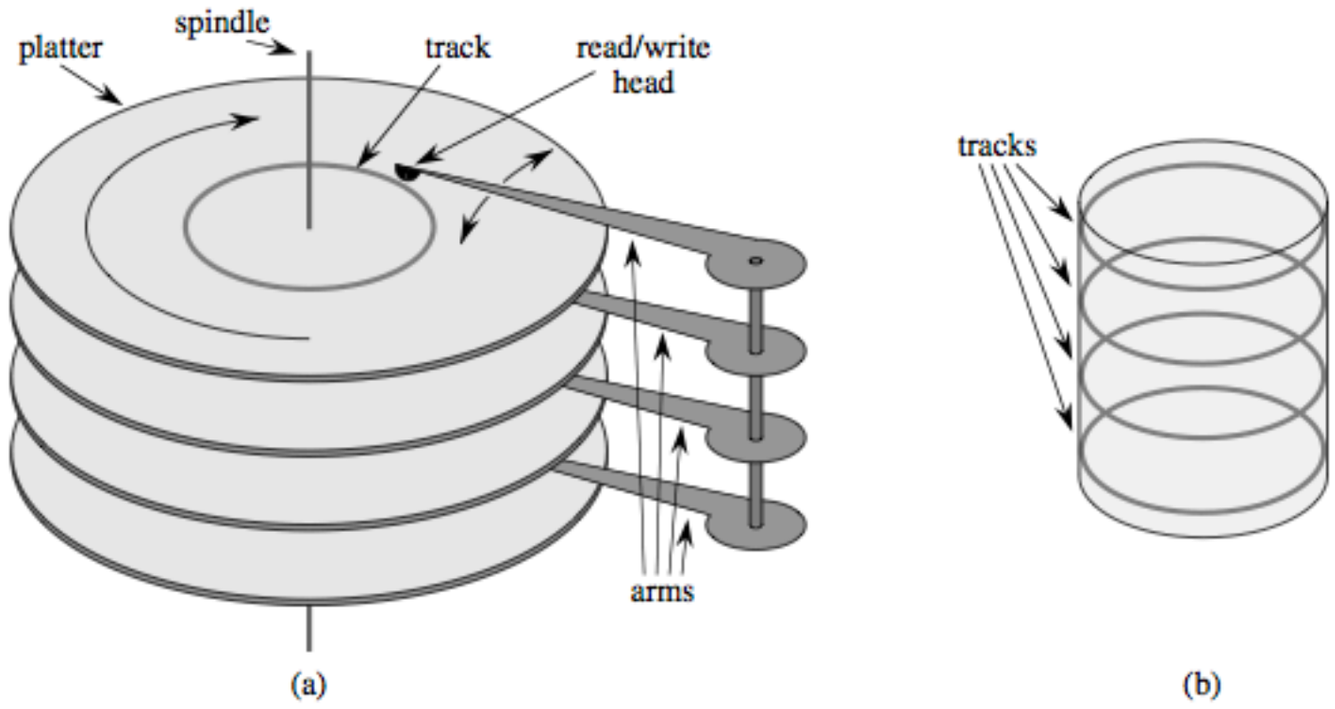


Fig. 2.1 Magnetic disk drive: (a) Data are stored on magnetized platters that rotate at a constant speed. Each platter surface is accessed by an arm that contains a read/write head, and data are stored on the platter in concentric circles called tracks. (b) The arms are physically connected so that they move in unison. The tracks (one per platter) that are addressable when the arms are in a fixed position are collectively referred to as a cylinder.

Striping

	$\mathcal{D}_0$	$\mathcal{D}_1$	$\mathcal{D}_2$	$\mathcal{D}_3$	$\mathcal{D}_4$
stripe 0	0 1	2 3	4 5	6 7	8 9
stripe 1	10 11	12 13	14 15	16 17	18 19
stripe 2	20 21	22 23	24 25	26 27	28 29
stripe 3	30 31	32 33	34 35	36 37	38 39

Parameters of IO-bound problem

N = problem size (in units of data items);

M = internal memory size (in units of data items);

B = block transfer size (in units of data items);

D = number of independent disk drives;

P = number of CPUs,

Q = number of queries (for a batched problem);

Z = answer size (in units of data items).

$$n = \frac{N}{B}, \quad m = \frac{M}{B}, \quad q = \frac{Q}{B}, \quad z = \frac{Z}{B}$$

Parallel Disk model of computation

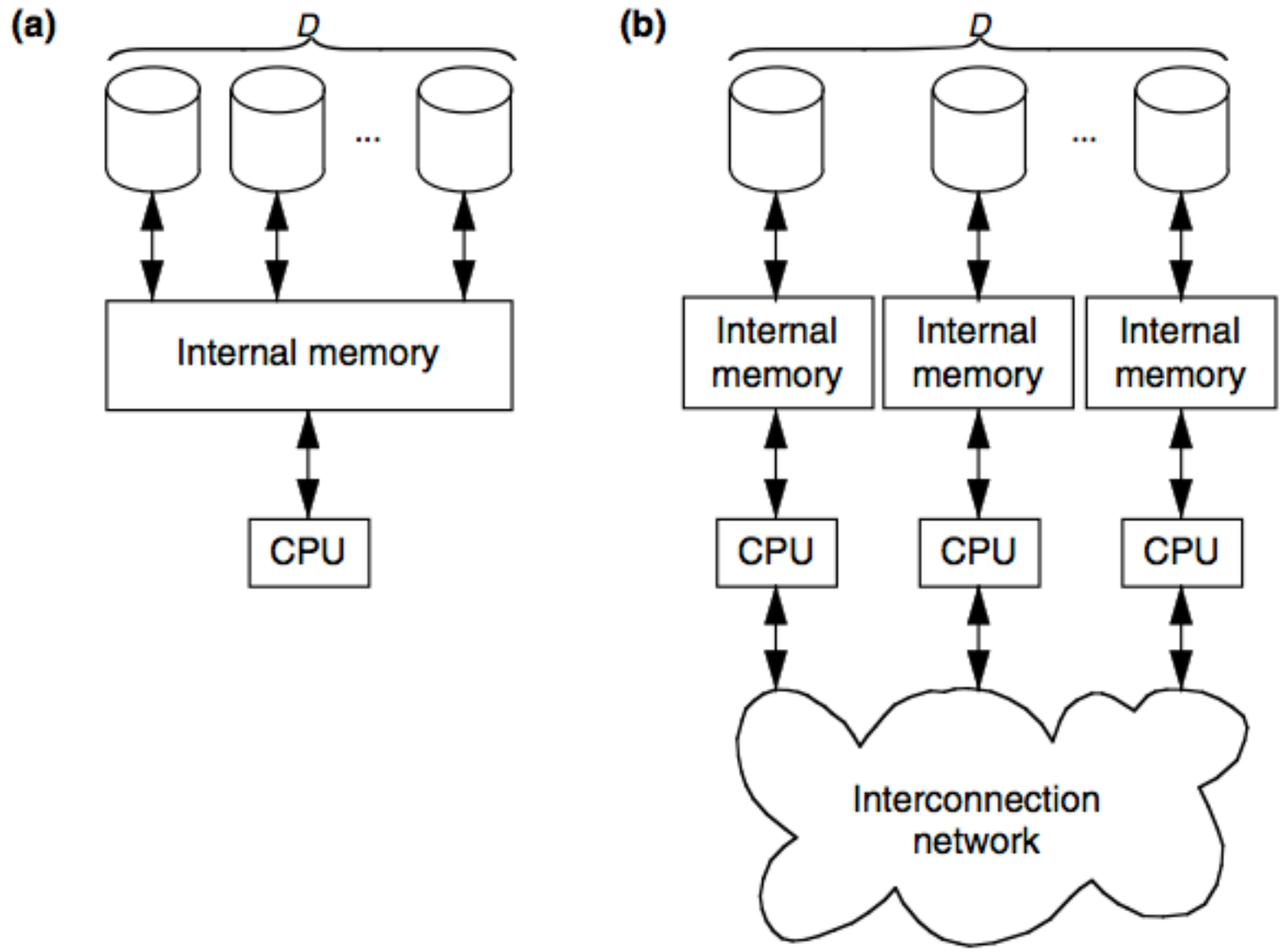


Fig. 2.2 Parallel disk model: (a) $P = 1$, in which the D disks are connected to a common CPU; (b) $P = D$, in which each of the D disks is connected to a separate processor.

Four Problems

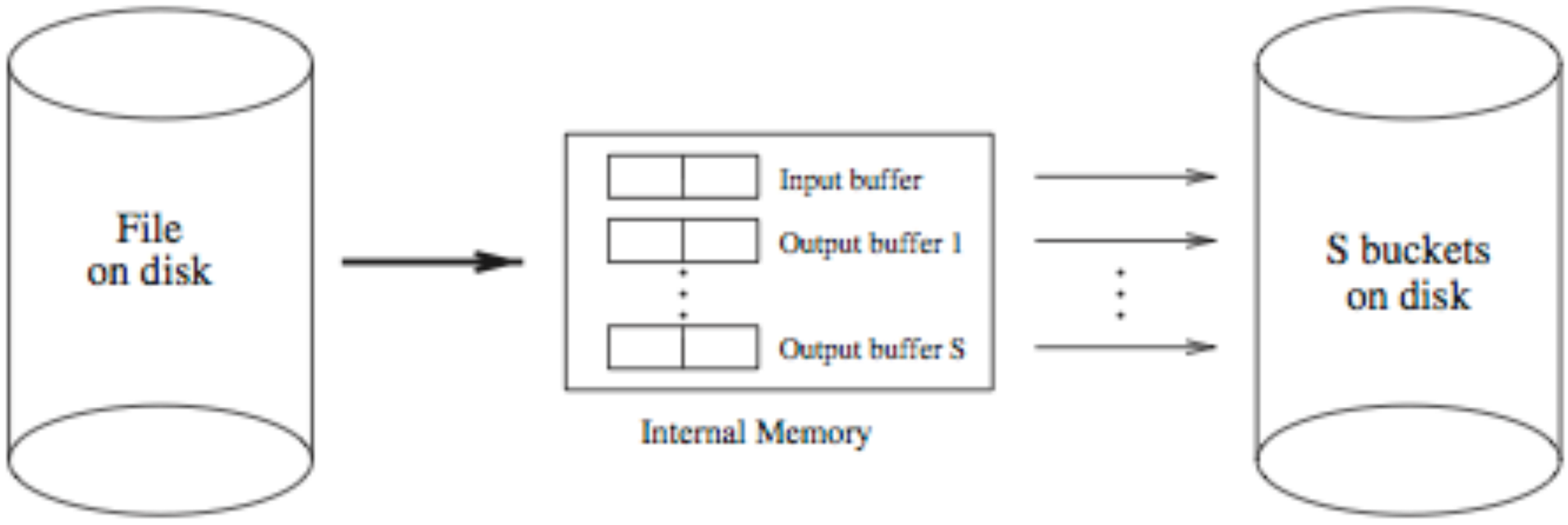
The I/O performance of many algorithms and data structures can be expressed in terms of the bounds for these fundamental operations:

(1) *Scanning* (a.k.a. *streaming* or *touching*) a file of N data items, which involves the sequential reading or writing of the items in the file.
(2) *Sorting* a file of N data items, which puts the items into sorted order.
(3) *Searching* online through N sorted data items.
(4) *Outputting* the Z items of an answer to a query in a blocked "output-sensitive" fashion.

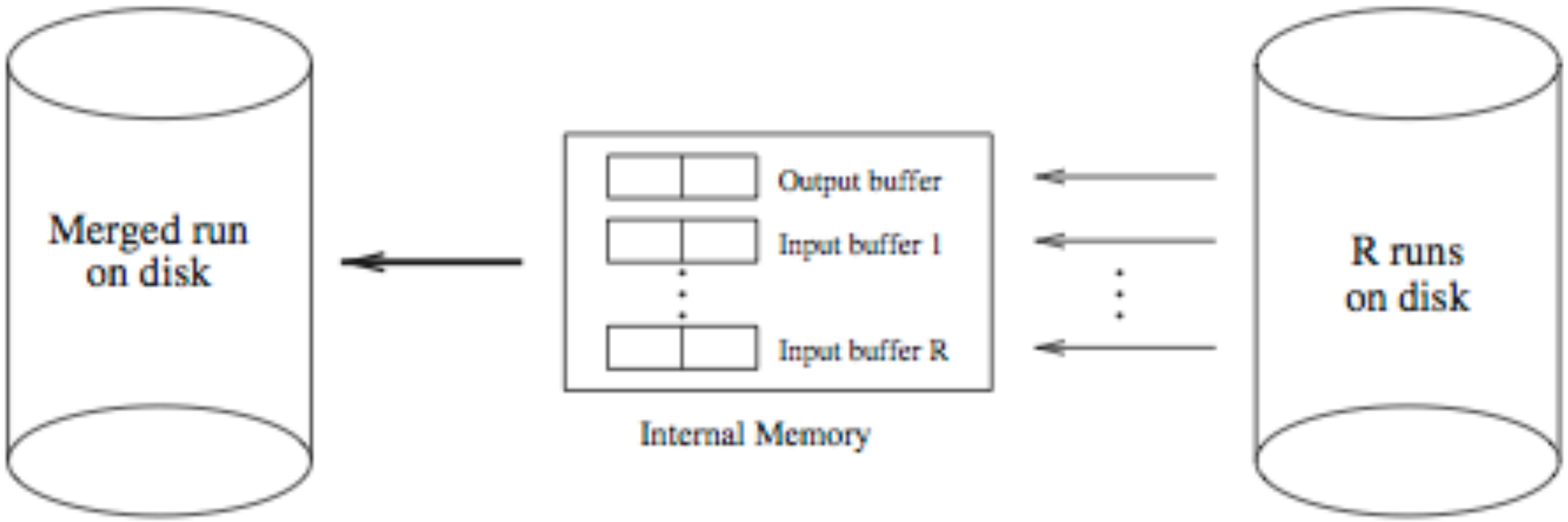
Performance bounds

Operation	I/O bound, $D = 1$	I/O bound, general $D \geq 1$
$Scan(N)$	$\Theta\left(\frac{N}{B}\right) = \Theta(n)$	$\Theta\left(\frac{N}{DB}\right) = \Theta\left(\frac{n}{D}\right)$
$Sort(N)$	$\Theta\left(\frac{N}{B}\log_{M/B}\frac{N}{B}\right)$ $= \Theta(n\log_m n)$	$\Theta\left(\frac{N}{DB}\log_{M/B}\frac{N}{B}\right)$ $= \Theta\left(\frac{n}{D}\log_m n\right)$
$Search(N)$	$\Theta(\log_B N)$	$\Theta(\log_{DB} N)$
$Output(Z)$	$\Theta\left(\max\left\{1, \frac{Z}{B}\right\}\right)$ $= \Theta(\max\{1, z\})$	$\Theta\left(\max\left\{1, \frac{Z}{DB}\right\}\right)$ $= \Theta\left(\max\left\{1, \frac{z}{D}\right\}\right)$

Bucket Sort


File
on disk
Input buffer
Output buffer 1
Output buffer S
Internal Memory
S buckets
on disk

Merge Sort


Merged run
on disk
Output buffer
Input buffer 1
Input buffer R
Internal Memory
R runs
on disk

More than one disk

- The challenge is to keep the result of each phase distributed uniformly across the disks.
- Combination of striping and randomization.