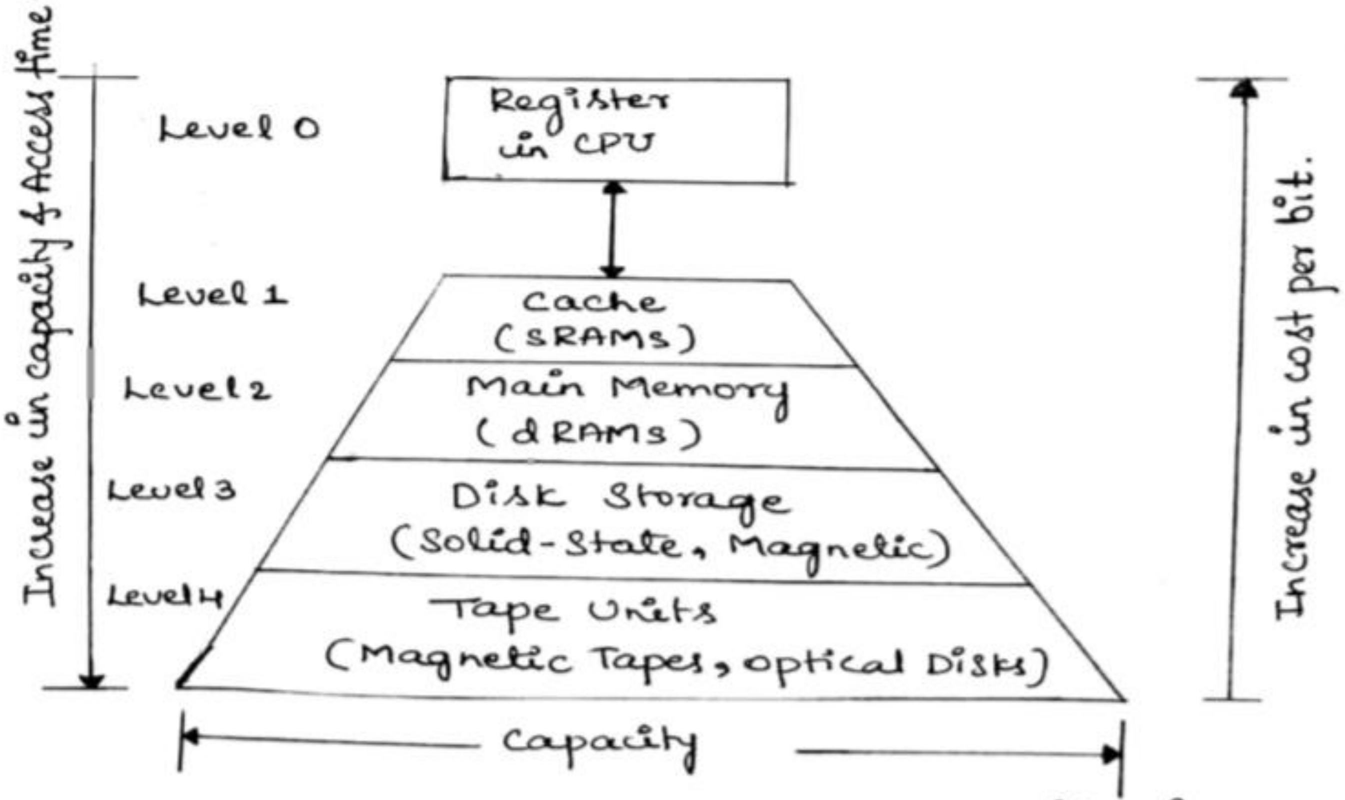
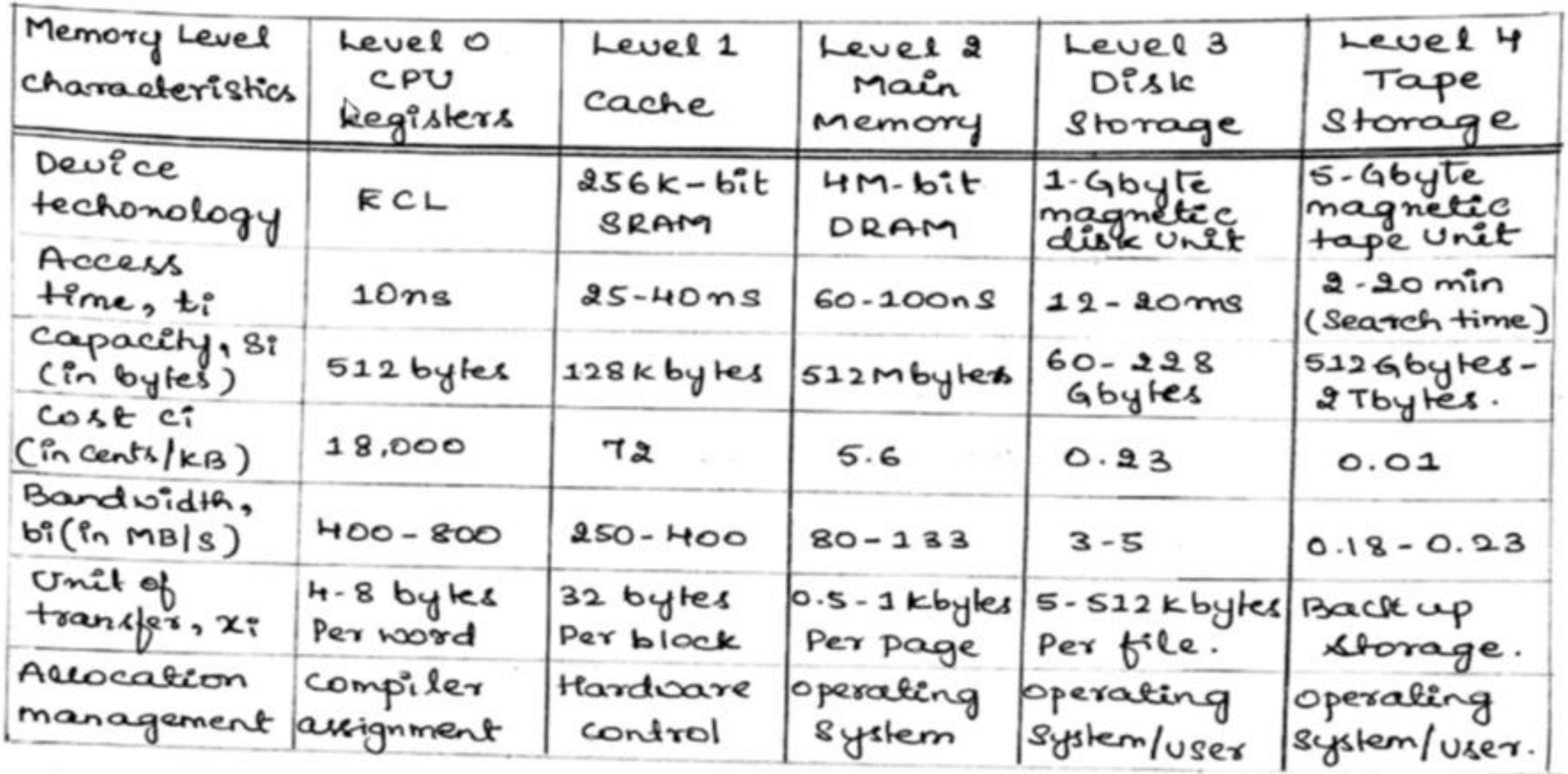
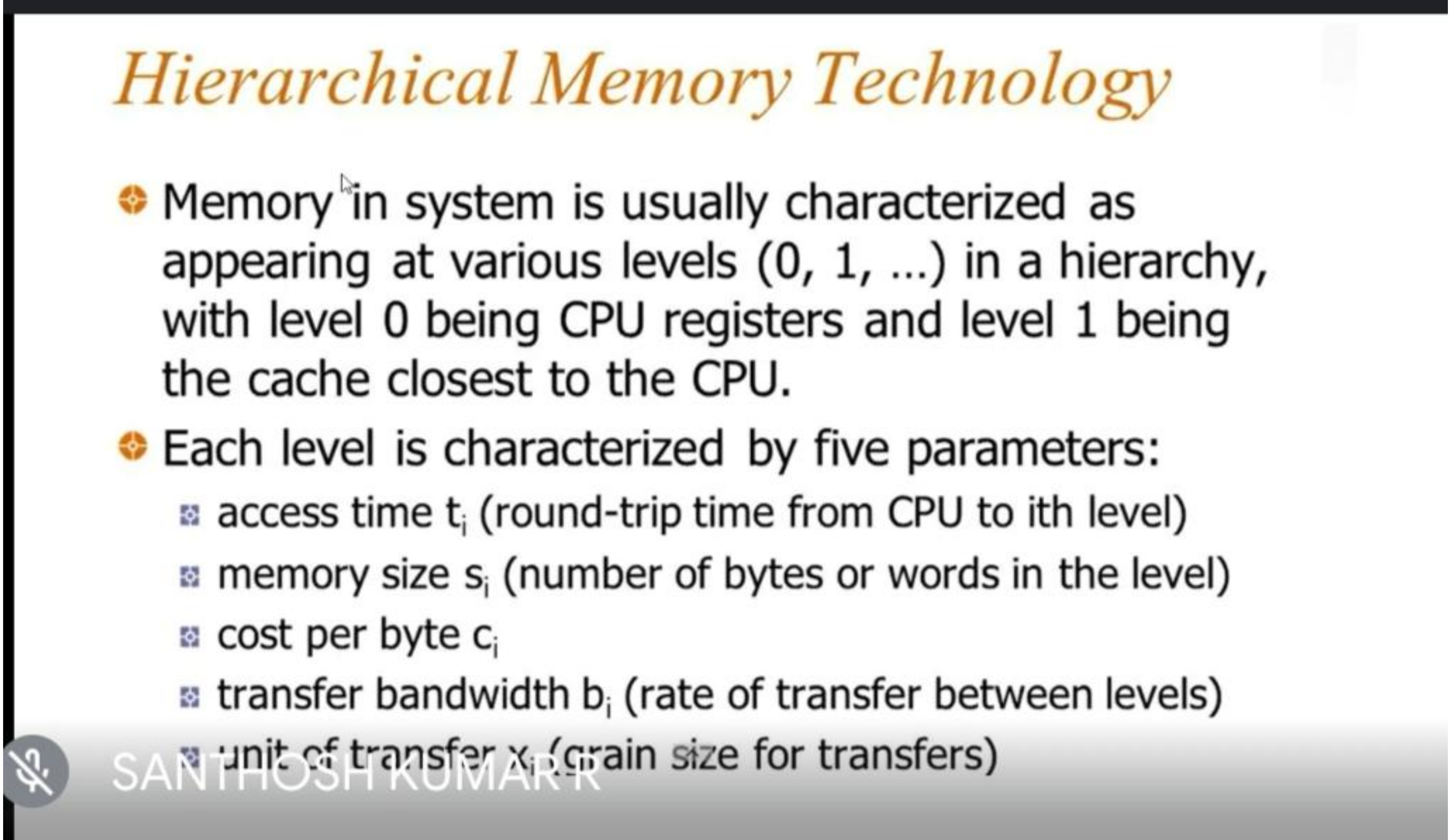

Increase in capacity & Access time
Level 0
Level 1
Level 2
Level 3
Level 4
Register in CPU
Cache (SRAMs)
Main Memory (dRAMs)
Disk Storage (Solid-State, Magnetic)
Tape Units (Magnetic Tapes, optical Disks)
capacity
Increase in cost per bit.

Figure : A four-level memory hierarchy with increasing capacity and decreasing speed and cost from low to high levels.



Memory Level characteristics	Level 0 CPU Registers	Level 1 Cache	Level 2 Main Memory	Level 3 Disk Storage	Level 4 Tape Storage
Device techonology	ECL	256K-bit SRAM	4M-bit DRAM	1-Gbyte magnetic disk unit	5-Gbyte magnetic tape unit
Access time, t_i	10ns	25-40ns	60-100ns	12-20ms	2-20 min (Search time)
Capacity, S_i (in bytes)	512 bytes	128K bytes	512M bytes	60-228 Gbytes	512 Gbytes-2 Tbytes.
Cost c_i (in cents/KB)	18,000	72	5.6	0.23	0.01
Bandwidth, b_i(in MB/s)	400-800	250-400	80-133	3-5	0.18-0.23
Unit of transfer, x_i	4-8 bytes Per word	32 bytes Per block	0.5-1 Kbytes Per page	5-512 Kbytes Per file.	Back up storage.
Allocation management	compiler assignment	Hardware control	operating system	operating system/user	operating system/user.

Table: Memory characteristics of a typical Mainframe computer in 1993



Hierarchical Memory Technology

- Memory in system is usually characterized as appearing at various levels (0, 1, …) in a hierarchy, with level 0 being CPU registers and level 1 being the cache closest to the CPU.
- Each level is characterized by five parameters:
  - access time t_i (round-trip time from CPU to ith level)
  - memory size s_i (number of bytes or words in the level)
  - cost per byte c_i
  - transfer bandwidth b_i (rate of transfer between levels)
  - unit of transfer x_i (grain size for transfers)

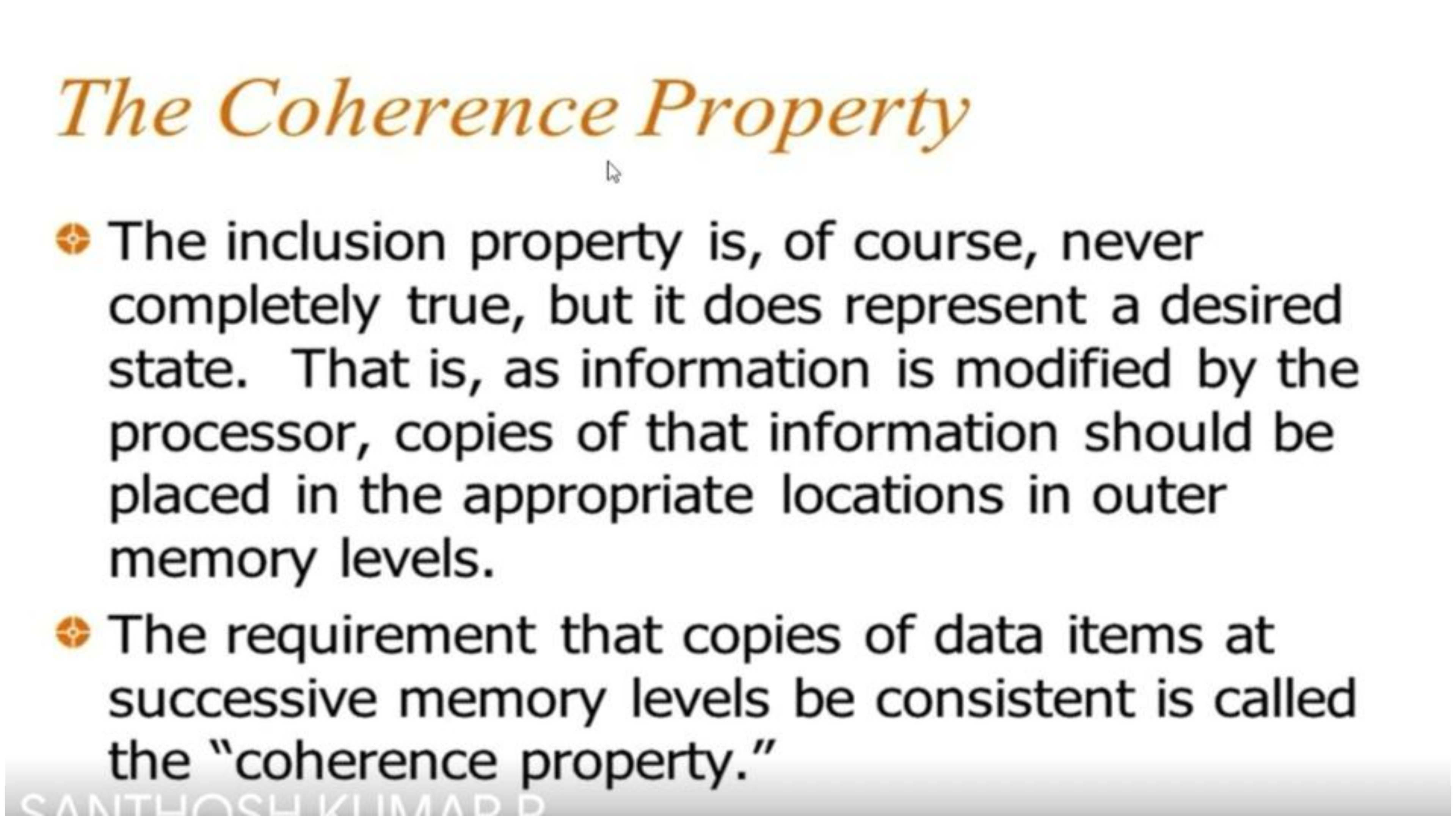
The Inclusion Property

- The inclusion property is stated as:

  $$M_1 \subset M_2 \subset \ldots \subset M_n$$

  The implication of the inclusion property is that all items of information in the "innermost" memory level (cache) also appear in the outer memory levels.

- The inverse, however, is not necessarily true. That is, the presence of a data item in level M_{i+1} does not imply its presence in level M_i. We call a reference to a missing item a "miss."

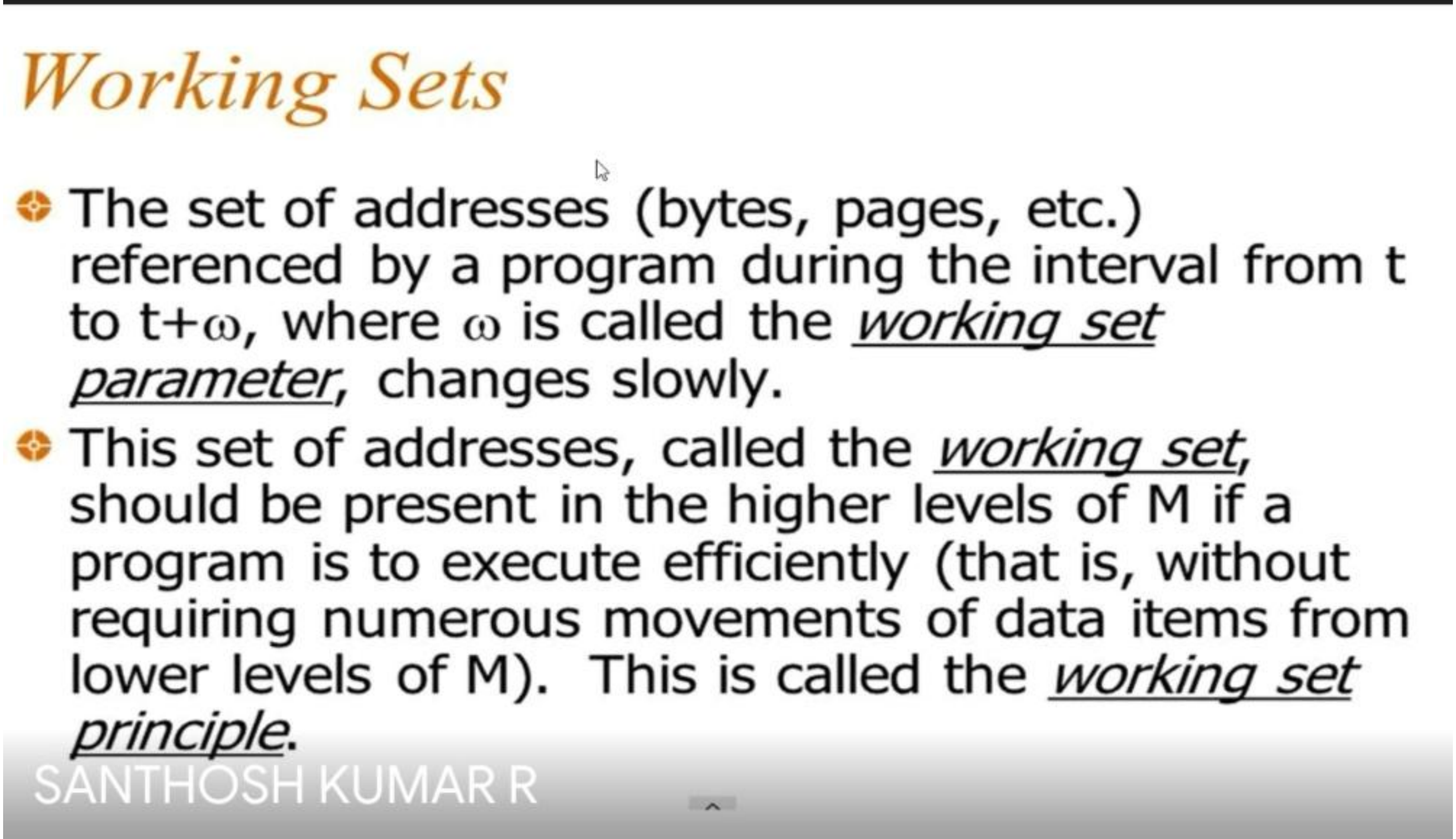


The Coherence Property

- The inclusion property is, of course, never completely true, but it does represent a desired state. That is, as information is modified by the processor, copies of that information should be placed in the appropriate locations in outer memory levels.
- The requirement that copies of data items at successive memory levels be consistent is called the "coherence property."

Locality of References

- In most programs, memory references are assumed to occur in patterns that are strongly related (statistically) to each of the following:
  - *Temporal locality* – if location M is referenced at time t, then it (location M) will be referenced again at some time t+Δt.
  - *Spatial locality* – if location M is referenced at time t, then another location M±Δm will be referenced at time t+Δt.
  - *Sequential locality* – if location M is referenced at time t, then locations M+1, M+2, … will be referenced at time t+Δt, t+Δt′, etc.
- In each of these patterns, both Δm and Δt are "small."
- H&P suggest that 90 percent of the execution time in most programs is spent executing only 10 percent of the code.



Working Sets

- The set of addresses (bytes, pages, etc.) referenced by a program during the interval from t to t+ω, where ω is called the *working set parameter*, changes slowly.
- This set of addresses, called the *working set*, should be present in the higher levels of M if a program is to execute efficiently (that is, without requiring numerous movements of data items from lower levels of M). This is called the *working set principle*.

Hit Ratios

- When a needed item (instruction or data) is found in the level of the memory hierarchy being examined, it is called a *hit*. Otherwise (when it is not found), it is called a *miss* (and the item must be obtained from a lower level in the hierarchy).
- The *hit ratio*, *h*, for M_i is the probability (between 0 and 1) that a needed data item is found when sought in level memory M_i.
- The *miss ratio* is obviously just $1-h_i$.
- We assume $h_0 = 0$ and $h_n = 1$.