The Memory Hierarchy

CS520
Dept. of Computer Science
Univ. of New Hampshire
Create illusion of very large memory that can be accessed with high speed.

Why illusion?

Because high-speed (i.e., low access time) memory is expensive.
<table>
<thead>
<tr>
<th>Technology</th>
<th>Access Time</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static RAM</td>
<td>1x</td>
<td>100x</td>
</tr>
<tr>
<td>Dynamic RAM</td>
<td>10x</td>
<td>1x</td>
</tr>
<tr>
<td>Disk</td>
<td>1Mx</td>
<td>0.01x</td>
</tr>
</tbody>
</table>
How can this possibly work?

Programs exhibit locality.

They access a relatively small portion of their address space in any small interval of time.
temporal locality

If a location is referenced, it will tend to be referenced again soon.
spatial locality

If a location is referenced, locations whose addresses are close by will tend to be referenced soon.
For \( i = 0 \) to \( i < N \) do

\[
\text{sum} = \text{sum} + a_{ij}.
\]
place locations in a hierarchy in which least frequently accessed locations are lowest in the hierarchy and most frequently accessed locations are highest in the hierarchy.
Spatial locality causes systems to treat blocks of contiguous memory locations as a single unit.
hit rate

Fraction of accesses satisfied at an upper level
miss penalty

time to satisfy access that misses at upper level
Key problems

determining where a location is in the hierarchy

moving locations up and down in the hierarchy