Larger, slower, cheaper storage device at level $k$ is partitioned into blocks.

Smaller, faster, more expensive device at level $k$ caches a subset of the blocks from level $k+1$.

Data are copied between levels in block-sized transfer units.