Lecture 13: DRAM Innovations

- Today: energy efficiency, row buffer management, scheduling

Latency and Power Wall

- Power wall: 25-40% of datacenter power can be attributed to the DRAM system

- Latency and power can be both improved by employing smaller arrays; incurs a penalty in density and cost

- Latency and power can be both improved by increasing the row buffer hit rate; requires intelligent mapping of data to rows, clever scheduling of requests, etc.

- Power can be reduced by minimizing overfetch – either read fewer chips or read parts of a row; incur penalties in area or bandwidth

Overfetch

- Overfetch caused by multiple factors:
  - Each array is large (fewer peripherals → more density)
  - Involving more chips per access → more data transfer pin bandwidth
  - More overfetch → more prefetch; helps apps with locality
  - Involving more chips per access → less data loss when a chip fails → lower overhead for reliability