Projects

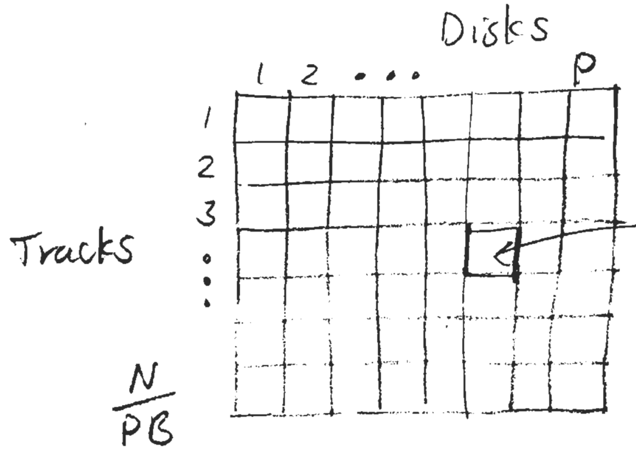
6.895
11/19/03
L20.1

Permuting data on parallel disks

Disk access times $\approx 10^{-2}$ sec
Data transfer rate $\approx 10^{-6}$ words/sec

$\therefore$ Want to do as few disk accesses as possible.

Convenient engineering assumption:
Disk is broken into large fixed-size <u>blocks</u>, e.g., of 1000 words.



P disks
N data records

Block = B data records

parallel IO's to read all data = N/PB.

Computer memory holds M data records total.
Assume $M \gg PB$.

Permuting disk blocks

- Off-line (perm fixed in advance)

← P disks →

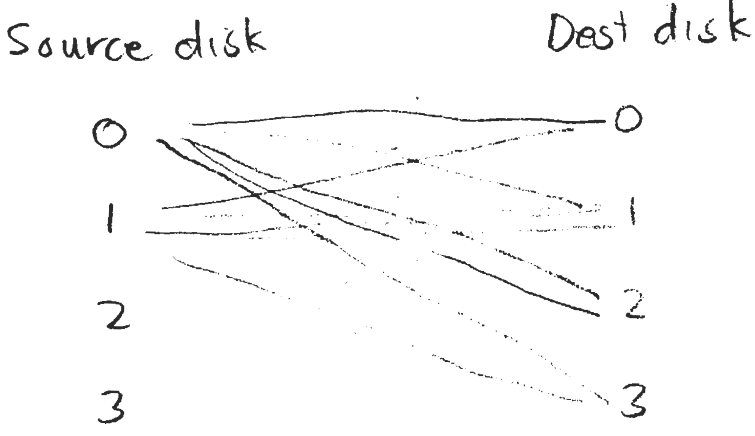
$n = N/P$ tracks

0	1	2	3
3,2 A	1,2 E	3,1 E	2,3 E
1,0 B	1,4 A	3,0 D	2,2 B
2,4 C	1,3 D	0,0 A	1,0 C
2,1 D	3,3 C	3,4 B	2,0 A
0,1 E	0,3 B	0,4 C	0,2 D

- $B = 1$

<u>Theorem</u>. Can permute with $O(N/P)$ parallel IO's (not in place)

Conflict graph

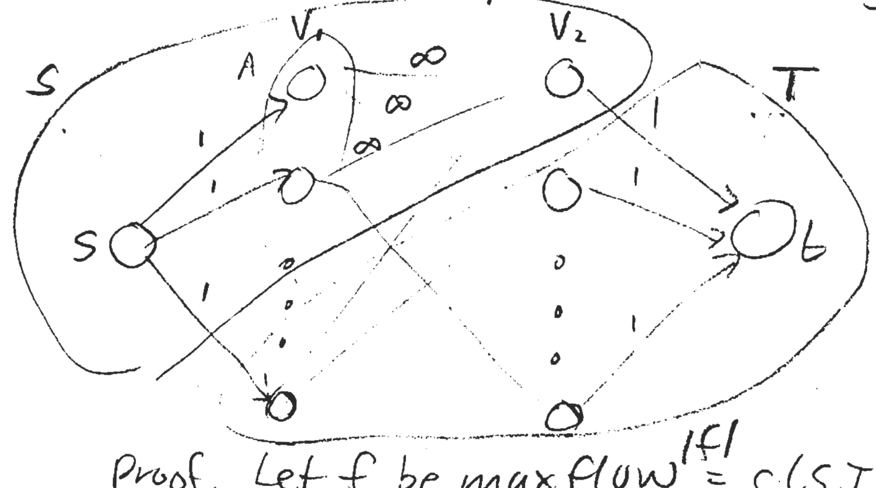


All degrees $= n = N/P$.

Fact: Any d-regular bipartite multigraph can be edge-colored with d colors. (Color = step at which block is moved.)

Method: Find perfect matching. Color edges in matching using color 1. Remove. Now have $(d-1)$-regular bipartite multigraph. Recur.

How do we know perfect matching exists?



Hall's Thm.
For $A \subseteq V_1$, let $N(A) \subseteq V_2$ be the set of neighbors of A. Then, a perfect matching exists iff $|N(A)| \geq |A|$ $\forall A$.

Proof. Let f be max flow. $|f| = c(S,T) < |V_1|$ for some cut (S,T) by maxflow-mincut thm.
Let $A = S \cap V_1$. Since edges from V_1 to V_2 have ∞ capacity, $N(A) \subseteq S$. Also, $N(V_1 - A) \subseteq T$.

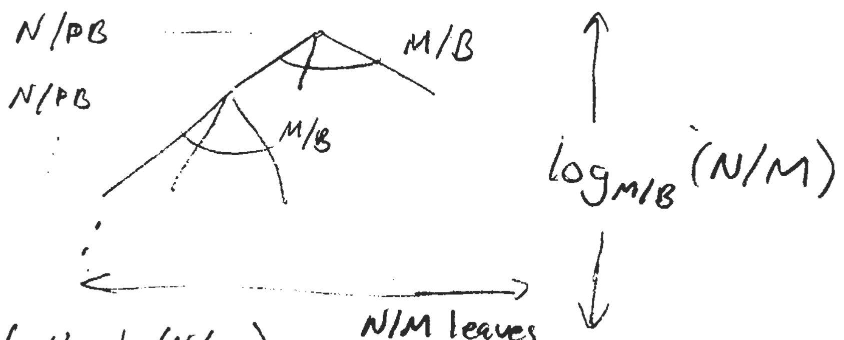
$$\therefore c(S,T) \geq |V_1 - A| + |N(A)|$$
$$\geq |V_1 - A| + |A|$$

Sorting (Vitter et al.)

$O\left(\frac{N}{PB} \cdot \frac{\lg(N/M)}{\lg(M/B)}\right)$ IO's.

Idea: Internal sort M records at a time into N/M runs.
Merge runs.

Would like to merge M/B runs at a time.



Total $\frac{N}{PB} \cdot \frac{\lg(N/M)}{\lg(M/B)}$

Problem: Can only read 1 block/run
- All of one run may be smaller than others.

Solution:
Merge $\sqrt{M/B}$ runs at a time (Depth of rec. doubled).
Keep track of which blocks to read next in table
"Sloppy" merge. Clean up with $O(N/PB)$ IO's.