Data Domain Deduplication - 2008 Reduce storage requirements for data archive - Improve performancé 1) Bloom filters (summary rector) 2) Statial locality (stream informed segment 3) locality preserved caching-prefetch sequent fingerprints together Beller beyering in system map object type ranges content store -> sequent descriptors segment store + index Ctingenprints) container manager objects - files/backup Affind good partitions image 2) shall hash Seg descriptor < > data 1) filtering diplicates locality 2) pack sequents into Loutsiner, when full give took 3) update segment in dex readingst container + orderbut portion, decompress, return segment on RAID-6 - writes fixed-sized containers in entirety

Techniques 1) Summery vector: avoid going to disk if data doesn't exist (not in index) Bloom titler-fake positives, not take megatives might have to check index sometimes K=4 hadrifunctions 2) Stream (uformed Layout Observation? sequente tend to occur together in same order leg. Dackup came file, or similar file) When were data contains deep segment X, high prob that other nearby segments here were near x in other 10 cales Dedicate container to wold sements in locical order (+ descriptors) logical order => Fewer 1/0s since get all e once 3) Locality Preserved Cading Traditional adding of Engerprists for index &
- vandom, not effective (high miss rate)

Cadre all fingerprints from container tradher
- Replace artire container

Steps on segment write

1) In segment conde? done

2) summany vector? no surew append to
current container

3) yes? check index

byes duplicate
insert into segment cocke

all fugurprints for this container

Cremere LEV

Best performance results? Most useful?

Table 4: Huge reduction in disk 1/0 W/ Bloom filter + locality preserved caching