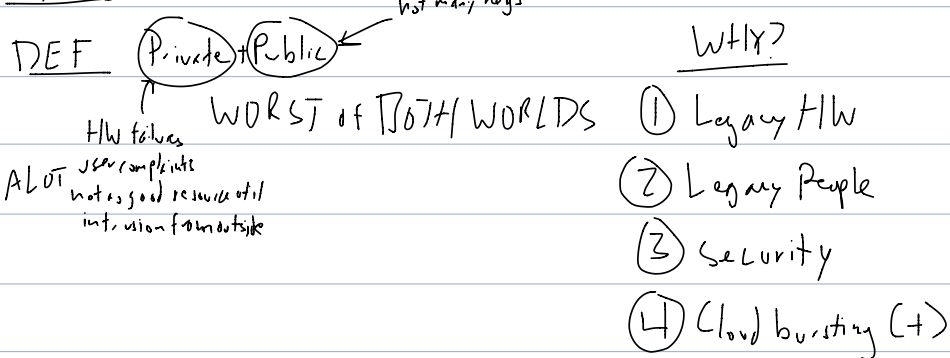


HYBRID CLOUD



"HYBRID" APP

- ① APP RUNS BOTH PLACES
- ② BOTH Private and Public Data

VMWARE + AWS

#1
private (b)
closed-source

	HDD	SSD	RAM
RANDOM IOPS	75-100	20K-100K	
MAX SIZE	8-16TB	1-2TB	
READ THROUGHPUT (RANDOM)			
READ THROUGHPUT (SEQ)	150MB/s	350MB/s	2-20 GB/s (20x SSD)

WRITE ^{THROUGHPUT}
(RANDOM)

WRITE ^{THROUGHPUT}
(SEQ)

LIFESPAN

ENERGY

RELIABILITY

HEAT

WEIGHT

LATENCY

PROGRAM:

1GB

COUNT # times each

HDD - $1GB / 150MB \approx 8 \text{ sec}$

Word appears in a doc

SSD -

1TB

$8000 \text{ sec} = 2 \text{ hrs}$

1TB - 1000 machines

Why not 8 seconds?

MERGE ← checkpoint

"WORD COUNT" "PAGE RANK" 2005 MAPREDUCE → Yahoo HADOOP

SUB1: tokenize input, route each token to a particular SUB2 (ALL SUB1's route by same rules)

SUB2: takes a stream of words, keeps running total of # times each word appears; at end, spit out list

IMPROVE

OPEN ISSUES

1. Parallelism (SPEED)

1. Machine Management (Death?)

2.

2. Distribution of input across machines

3. Cost to bandwidth