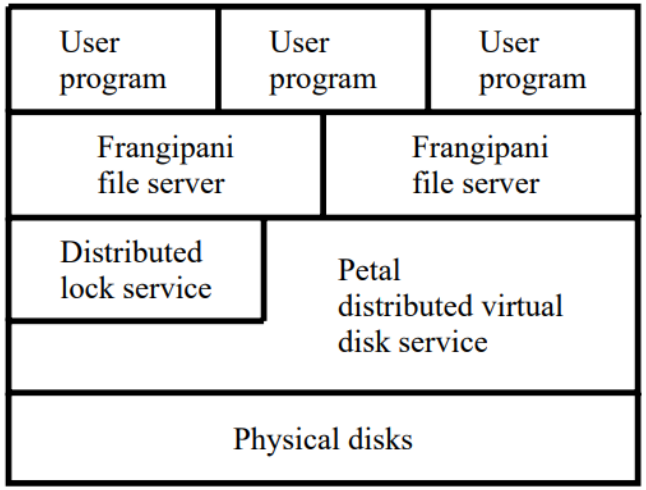
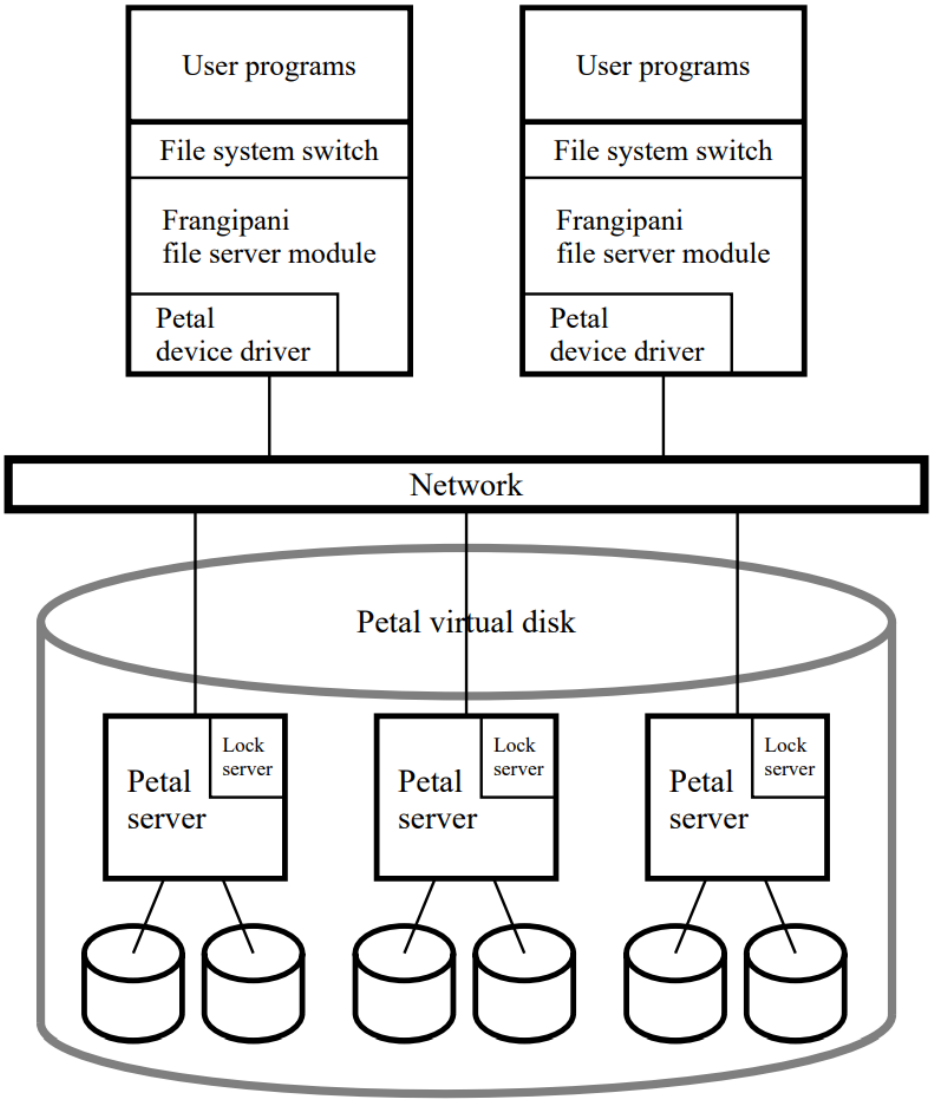
ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

HY590.45
Modern Topics in Scalable Storage Systems

Kostas Magoutis
magoutis@csd.uoc.gr
http://www.csd.uoc.gr/~hy590-45

Frangipani: Layering and structure

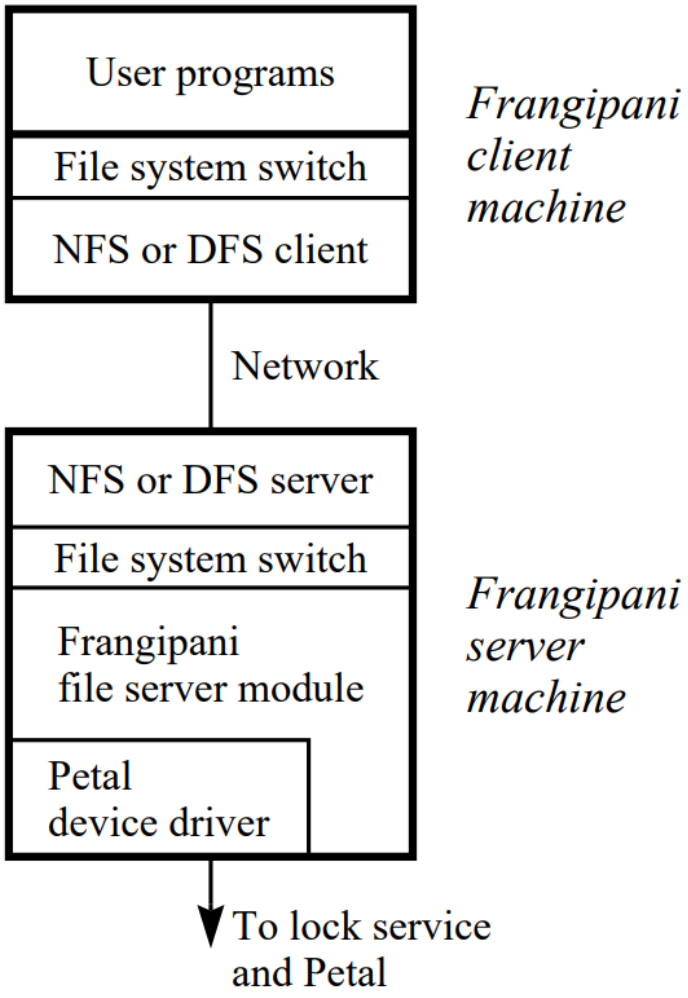




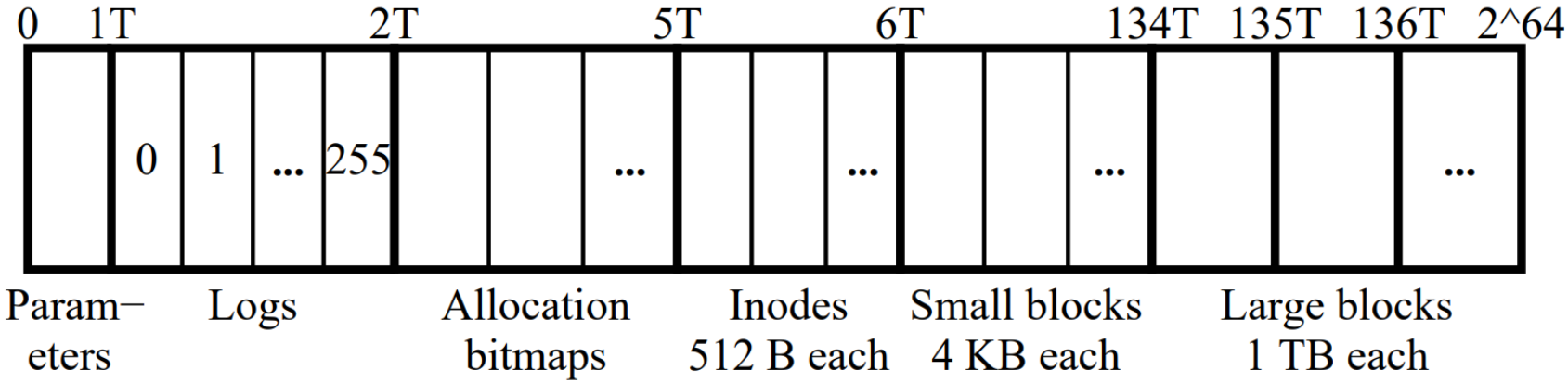
Goals

- Coherent, shared access to files across many users
- Ability to scalably add more servers, disk capacity
- Add new users without worrying about management
- Full, online, consistent backup of entire file systems
- Ability to tolerate machine, network, disk failures

Client/server configuration



Disk layout



- Example: Lookup file /dir/file1
- Example: Create file /dir/file2

Metadata logging

- Each metadata update is recorded in the log
- Logs are periodically written out to Petal
- In-place update performed by updated, every 30sec
- Logs are 128KB
- Recovery is run on a log after failure is detected

Locking

- Multiple readers/single writer
  - Sticky locks
  - Using leases (30”)
- Four basic operations
- Three implementations
  - Single lock server with volatile state
  - Primary/backup lock server, state stored on Petal
  - Distributed lock server with volatile state
- Global state replicated across all lock servers

Recovery

- Recovery daemon is given ownership of log and locks
- Runs log, releases locks
- Important issues:
  - Serialize updates to same data by different servers
  - Apply only updates that were logged since the server acquired the locks that cover them and for which it still holds the locks

Design issues

- Logging happens twice
- Cannot use disk location info for placing data
- Locking entire files, not individual blocks

Modified Andrew benchmark

Phase	Description	AdvFS		Frangipani	
		Raw	NVR	Raw	NVR
1	Create Directories	0.69	0.66	0.52	0.51
2	Copy Files	4.3	4.3	5.8	4.6
3	Directory Status	4.7	4.4	2.6	2.5
4	Scan Files	4.8	4.8	3.0	2.8
5	Compile	27.8	27.7	31.8	27.8

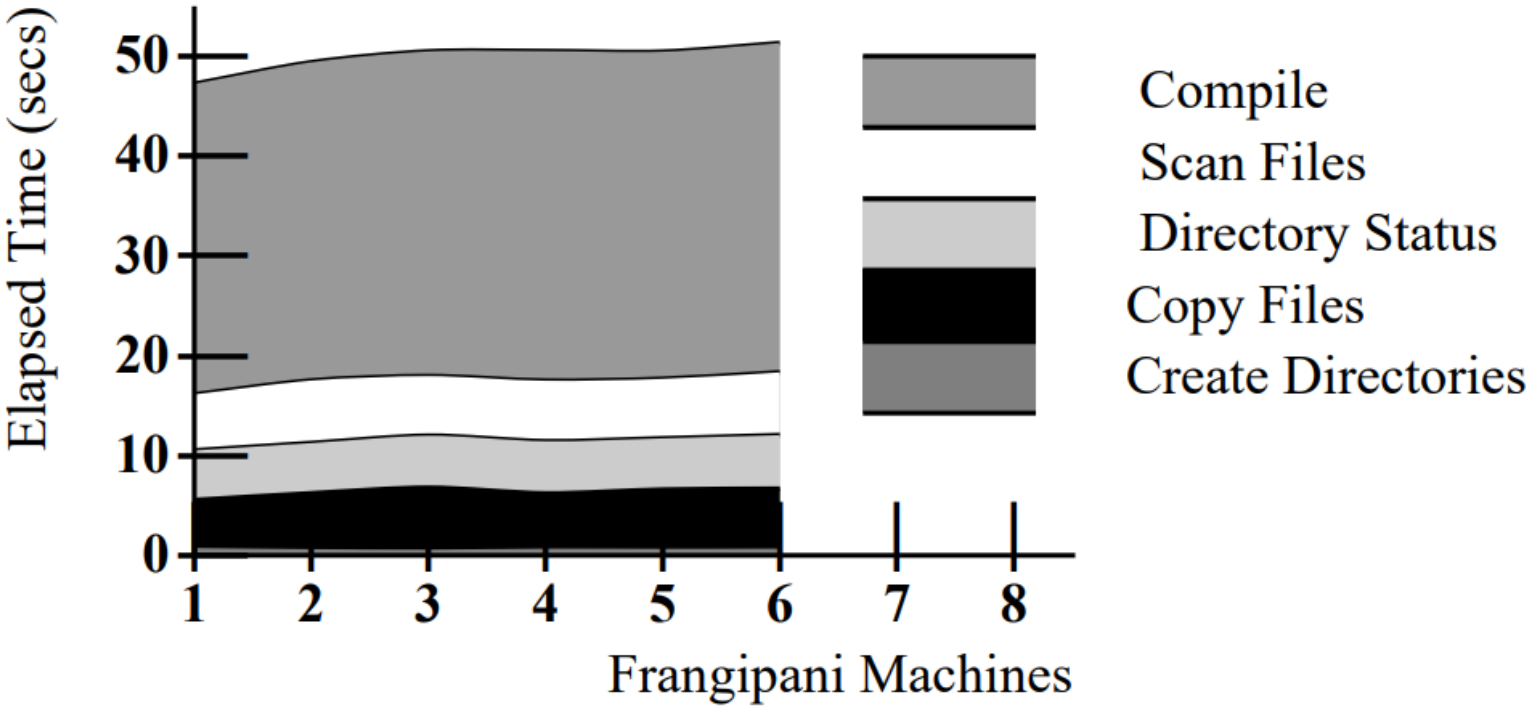
Connectathon benchmark

Test	Description	AdvFS		Frangipani	
		Raw	NVR	Raw	NVR
1	file and directory creation: creates 155 files and 62 directories.	0.92	0.80	3.11	2.37
2	file and directory removal: removes 155 files and 62 62 directories.	0.62	0.62	0.43	0.43
3	lookup across mount point: 500 getwd and stat calls.	0.56	0.56	0.43	0.40
4	setattr, getattr, and lookup: 1000 chmods and stats on 10 files.	0.42	0.40	1.33	0.68
5a	write: writes a 1048576 byte file 10 times.	2.20	2.16	2.59	1.63
5b	read: reads a 1048576 byte file 10 times.	0.54	0.45	1.81	1.83
6	readdir: reads 20500 directory entries, 200 files.	0.58	0.58	2.63	2.34
7	link and rename: 200 renames and links on 10 files.	0.47	0.44	0.60	0.50
8	symlink and readlink: 400 symlinks and readlinks on 10 files.	0.93	0.82	0.52	0.50
9	statfs: 1500 statfs calls.	0.53	0.49	0.23	0.22

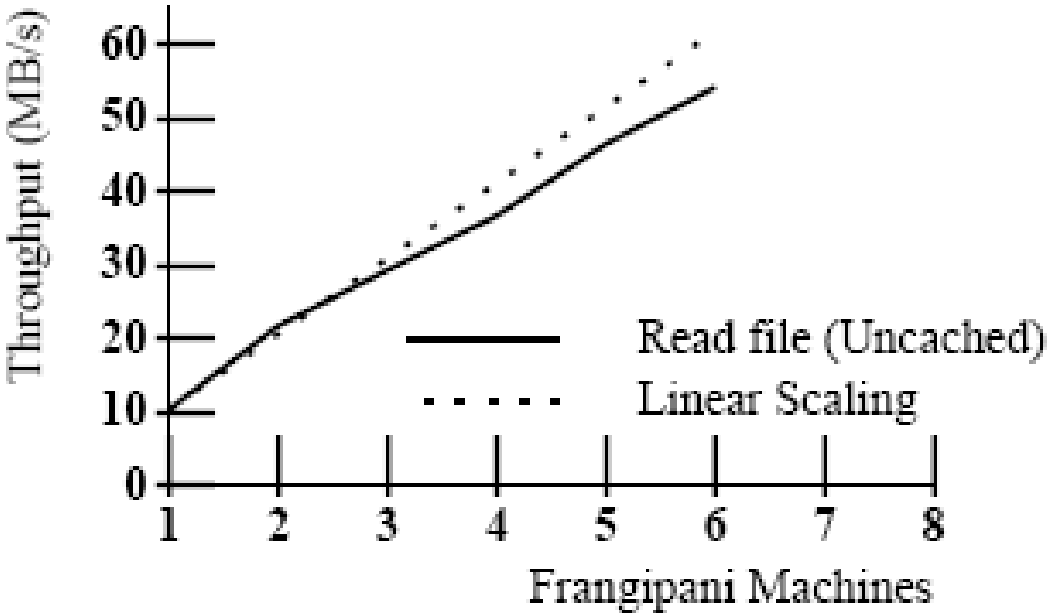
Throughput and CPU utilization

	Throughput (MB/s)		CPU Utilization	
	Frangipani	**AdvFS**	**Frangipani**	**AdvFS**
Write	15.3	13.3	42%	80%
Read	10.3	13.2	25%	50%

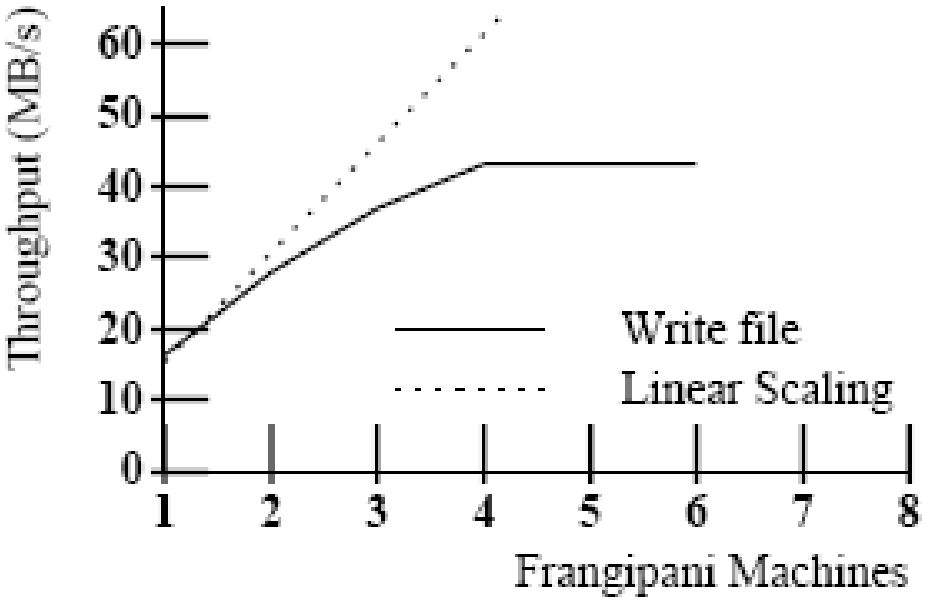
Scaling on modified Andrew



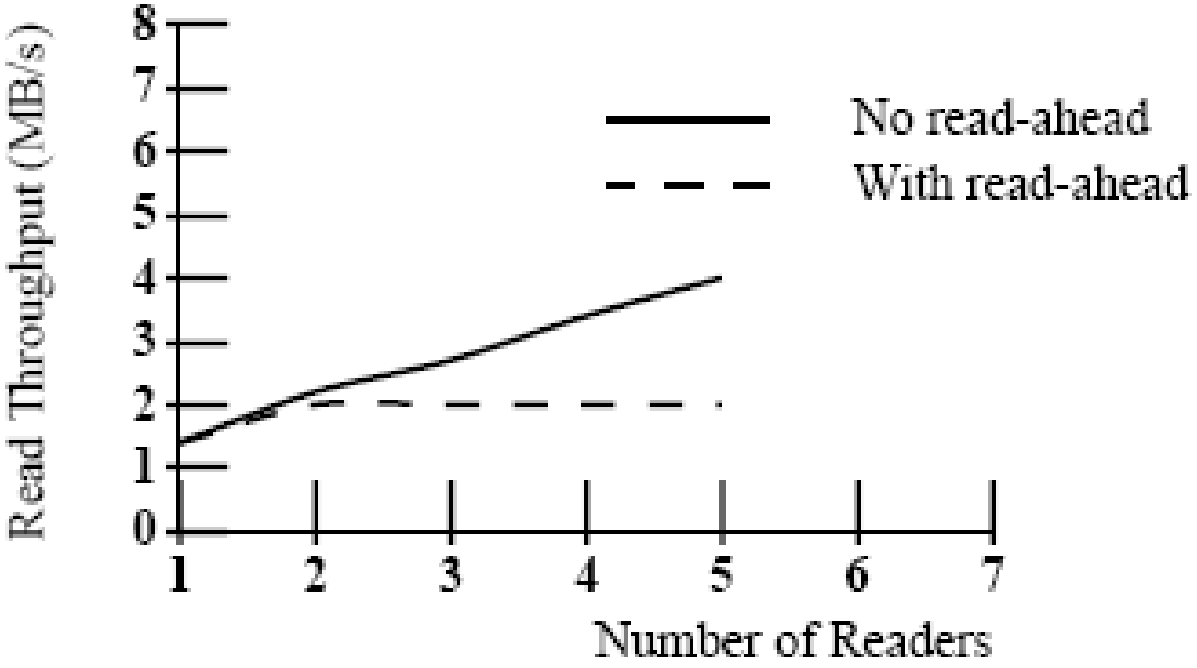
Scaling on uncached read



Scaling on writes



Reader/writer contention



Effect of size on reader/writer contention