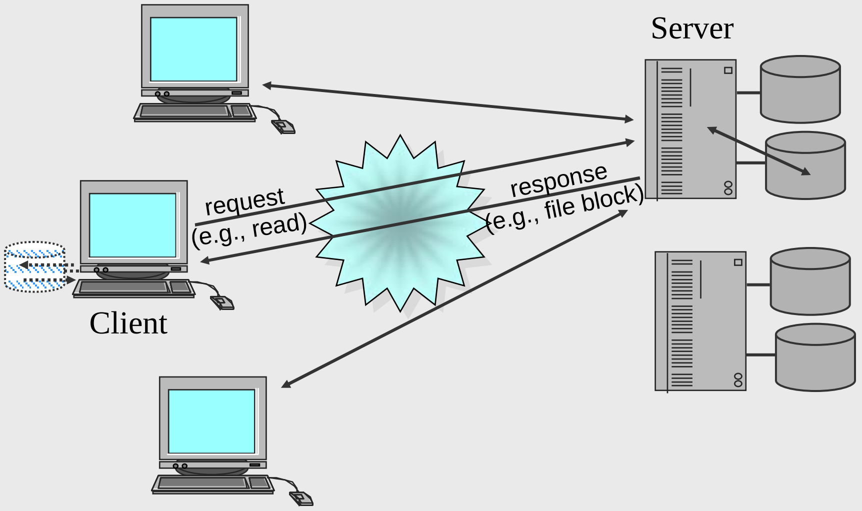
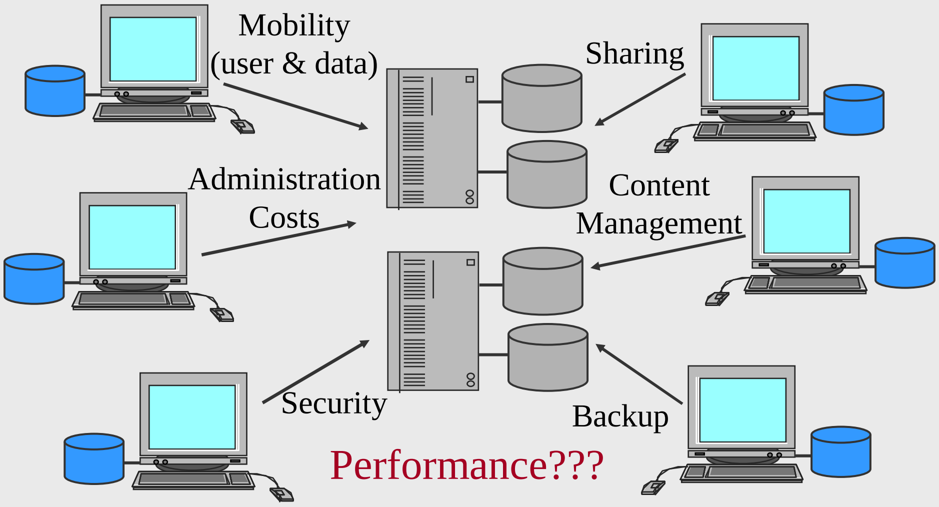
COMP 790-088 -- Distributed File Systems

With Case Studies:
Andrew and Google

File System Client and Server



Factors Encouraging Migration of Data to Shared File Systems



Mobility (user & data)

Sharing

Administration Costs

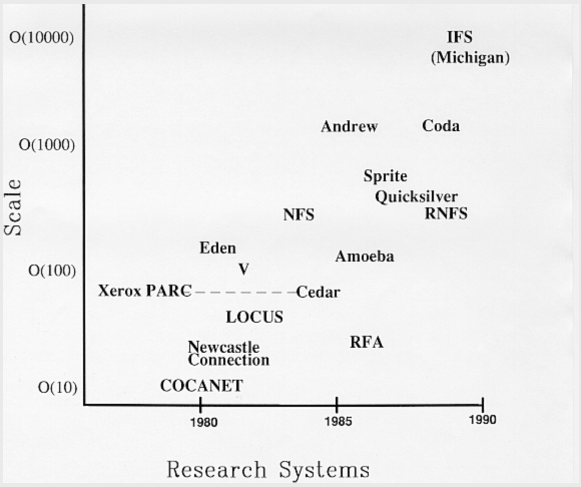
Content Management

Security

Backup

Performance???

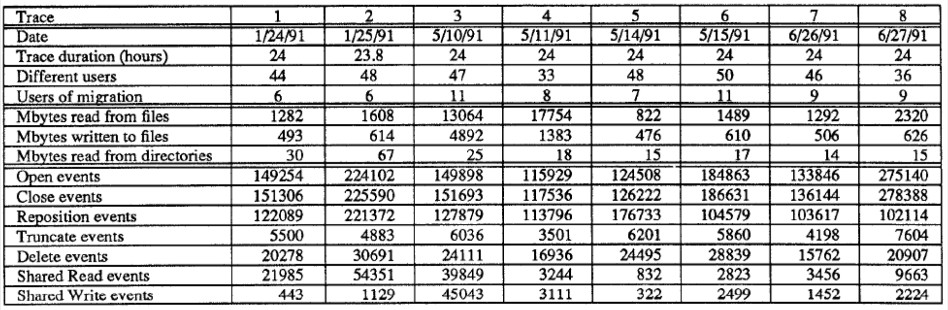
Chronology of Early File Systems



Research Systems

Summary of Sprite Study (1991)

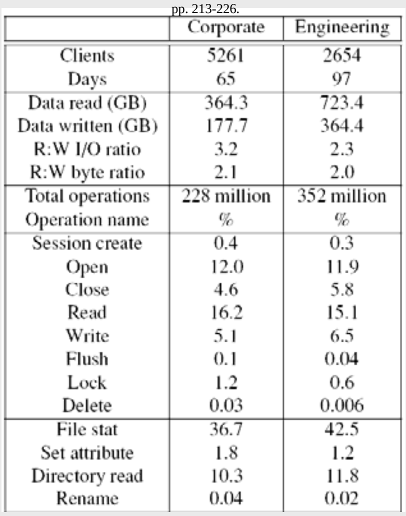
Source: Mary Baker, et at, “Measurements of a Distributed File System,” Proceedings 13th ACM SOSP, 1991, pp. 198-212.



Trace	1	2	3	4	5	6	7	8
Date	1/24/91	1/25/91	5/10/91	5/11/91	5/14/91	5/15/91	6/26/91	6/27/91
Trace duration (hours)	24	23.8	24	24	24	24	24	24
Different users	44	48	47	33	48	50	46	36
Users of migration	6	6	11	8	7	11	9	9
Mbytes read from files	1282	1608	13064	17754	822	1489	1292	2320
Mbytes written to files	493	614	4892	1383	476	610	506	626
Mbytes read from directories	30	67	25	18	15	17	14	15
Open events	149254	224102	149898	115929	124508	184863	133846	275140
Close events	151306	225590	151693	117536	126222	186631	136144	278388
Reposition events	122089	221372	127879	113796	176733	104579	103617	102114
Truncate events	5500	4883	6036	3501	6201	5860	4198	7604
Delete events	20278	30691	24111	16936	24495	28839	15762	20907
Shared Read events	21985	54351	39849	3244	832	2823	3456	9663
Shared Write events	443	1129	45043	3111	322	2499	1452	2224

Summary of NetApp Study (2008)

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.



	Corporate	Engineering
Clients	5261	2654
Days	65	97
Data read (GB)	364.3	723.4
Data written (GB)	177.7	364.4
R:W I/O ratio	3.2	2.3
R:W byte ratio	2.1	2.0
Total operations	228 million	352 million
Operation name	%	%
Session create	0.4	0.3
Open	12.0	11.9
Close	4.6	5.8
Read	16.2	15.1
Write	5.1	6.5
Flush	0.1	0.04
Lock	1.2	0.6
Delete	0.03	0.006
File stat	36.7	42.5
Set attribute	1.8	1.2
Directory read	10.3	11.8
Rename	0.04	0.02

Comparison of Studies

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.

File System Type	2008		Network		2003		1991	2000	Local	1999
Workload	Corporate		Engineering		CAMPUS	EECS	Sprite	Ins	Res	NT
Access Pattern	I/Os	Bytes	I/Os	Bytes	Bytes	Bytes	Bytes	Bytes	Bytes	Bytes
Read-Only (% total)	39.0	52.1	50.6	55.3	53.1	16.6	83.5	98.7	91.0	59.0
Entire file sequential	13.5	10.5	35.2	27.4	47.7	53.9	72.5	86.3	53.0	68.0
Partial sequential	58.4	69.2	45.0	55.0	29.3	36.8	25.4	5.9	23.2	20.0
Random	28.1	20.3	19.8	17.6	23.0	9.3	2.1	7.8	23.8	12.0
Write-Only (% total)	15.1	25.2	17.3	23.6	43.8	82.3	15.4	1.1	2.9	26.0
Entire file sequential	21.2	36.2	15.6	35.2	37.2	19.6	67.0	84.7	81.0	78.0
Partial sequential	57.6	55.1	63.4	61.0	52.3	76.2	28.9	9.3	16.5	7.0
Random	21.2	8.7	21.0	3.8	10.5	4.1	4.0	6.0	2.5	15.0
Read-Write (% total)	45.9	22.7	32.1	21.1	3.1	1.1	1.1	0.2	6.1	15.0
Entire file sequential	7.4	0.1	0.4	0.1	1.4	4.4	0.1	0.1	0.0	22.0
Partial sequential	48.1	78.3	27.5	50.0	0.9	1.8	0.0	0.2	0.3	3.0
Random	44.5	21.6	72.1	49.9	97.8	93.9	99.9	99.6	99.7	74.0

Windows

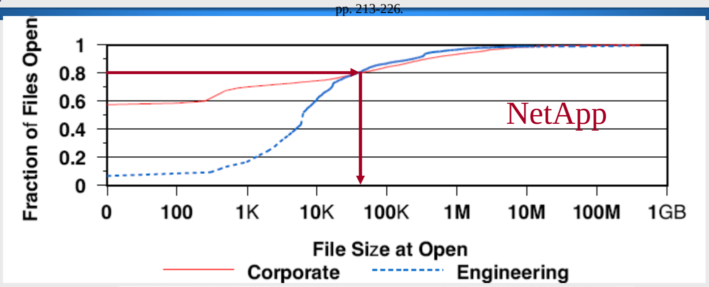
Unix

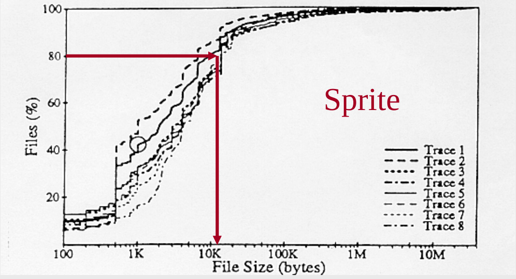
Windows

File Sizes (by % Files Accessed)

Source: Mary Baker, et at, “Measurements of a Distributed File System,” Proceedings 13th ACM SOSP, 1991, pp. 198-212.

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.



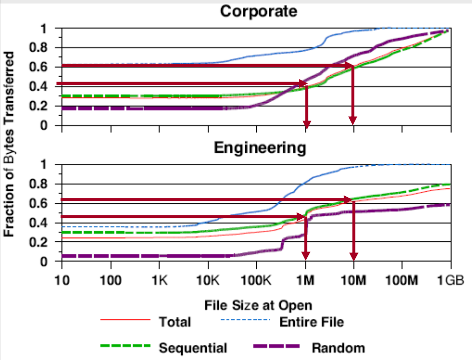


File Sizes (by % Bytes Transferred)

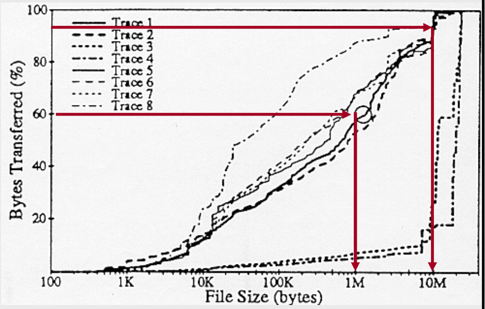
Source: Mary Baker, et at, “Measurements of a Distributed File System,” Proceedings 13th ACM SOSP, 1991, pp. 198-212.

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008 pp. 213-226.

NetApp



Sprite

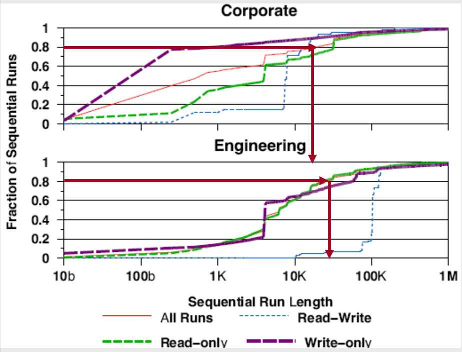


Run Length (by % Runs)

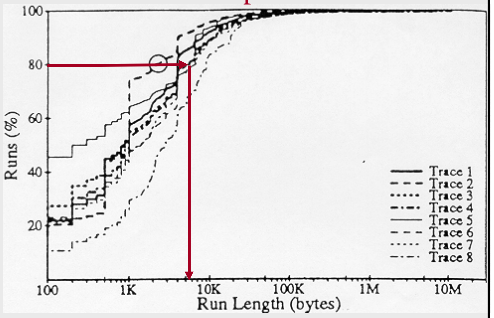
Source: Mary Baker, et at, “Measurements of a Distributed File System,” Proceedings 13th ACM SOSP, 1991, pp. 198-212.

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.

NetApp



Sprite

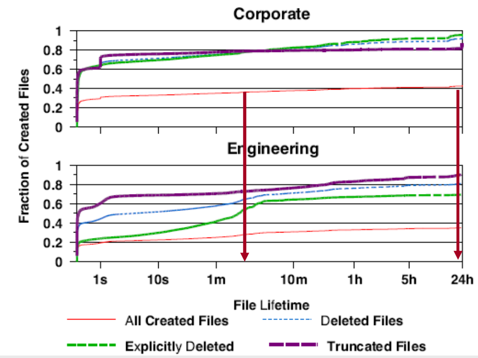


File Lifetimes (by % Files)

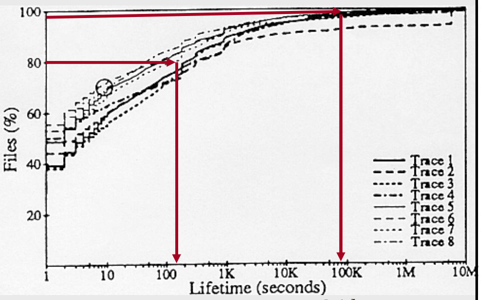
Source: Mary Baker, et at, “Measurements of a Distributed File System,” Proceedings 13th ACM SOSP, 1991, pp. 198-212.

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.

NetApp



Sprite



2 min

24 hrs

Modes of Sharing a Single File

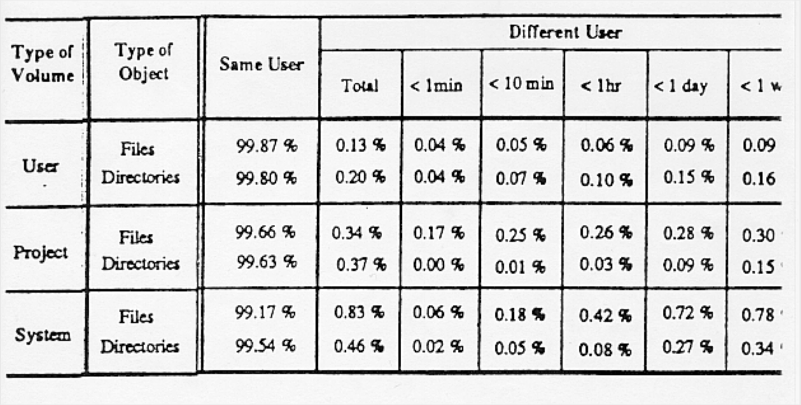
- Sequential Read Sharing
  - two or more read operations *do not* overlap in time
- Sequential Write Sharing
  - two or more operations, at least one of which is a write, *do not* overlap in time
- Concurrent Read Sharing
  - two or more read operations overlap in time
- Concurrent Write Sharing
  - two or more operations, at least one of which is a write, overlap in time

Strong Semantics for Concurrent Write Sharing

- Writes from multiple writers are “atomic”
  - subsequent reader sees entire update from one of the writers, never some partial update or merging of multiple updates
- Readers always see the atomic result of the most recently completed write operation

Statistics of File Sharing (Unix)

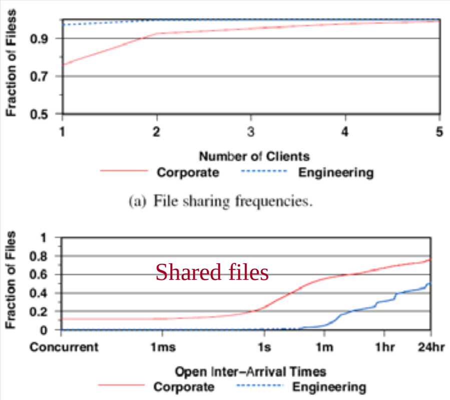
Source: Kistler and Satyanarayanan, “Disconnected Operation in the Coda File System, ACM TOCS, vol. 10, no. 1, Feb. 1992.



Type of Volume	Type of Object	Same User	Different User					
			Total	< 1min	< 10 min	< 1hr	< 1 day	< 1 w
User	Files	99.87 %	0.13 %	0.04 %	0.05 %	0.06 %	0.09 %	0.09
	Directories	99.80 %	0.20 %	0.04 %	0.07 %	0.10 %	0.15 %	0.16
Project	Files	99.66 %	0.34 %	0.17 %	0.25 %	0.26 %	0.28 %	0.30
	Directories	99.63 %	0.37 %	0.00 %	0.01 %	0.03 %	0.09 %	0.15
System	Files	99.17 %	0.83 %	0.06 %	0.18 %	0.42 %	0.72 %	0.78
	Directories	99.54 %	0.46 %	0.02 %	0.05 %	0.08 %	0.27 %	0.34

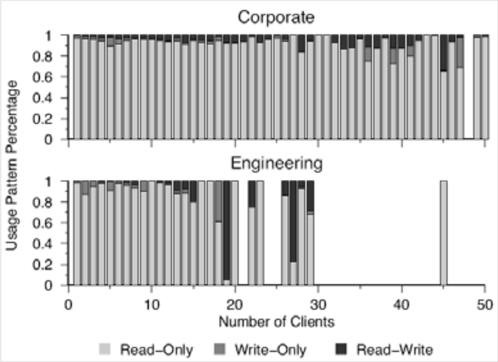
Statistics of File Sharing (Windows)

Source: Andrew W. Leung, et at, “Measurement and Analysis of Large-Scale Network File System Workloads,” Proceedings USENIX Annual Technical Conference, 2008, pp. 213-226.



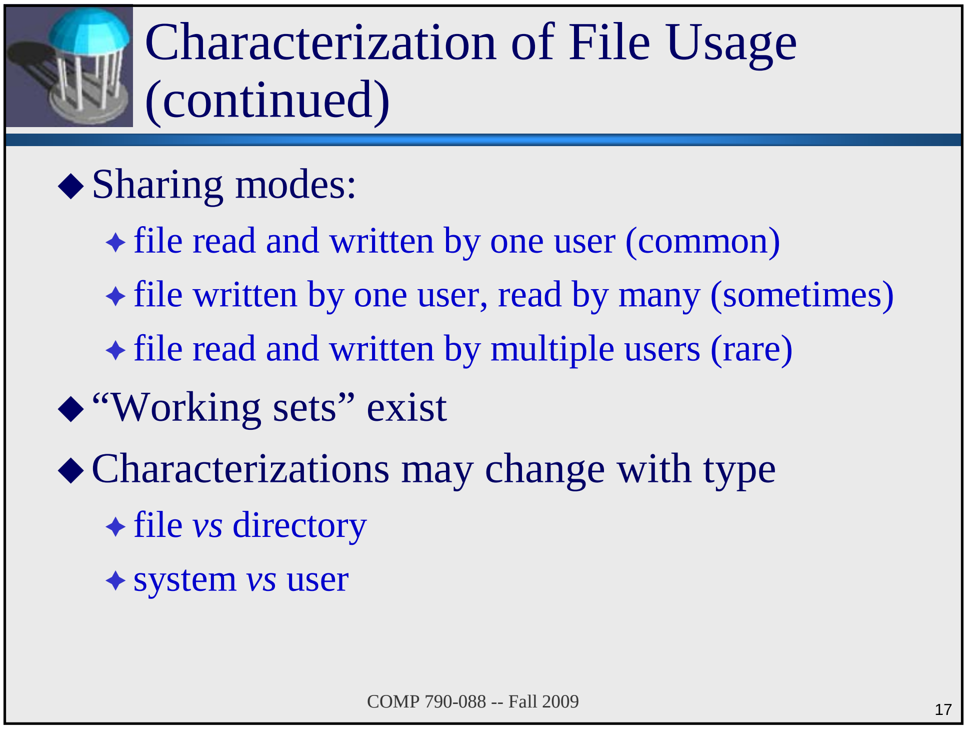
(a) File sharing frequencies.

Shared files



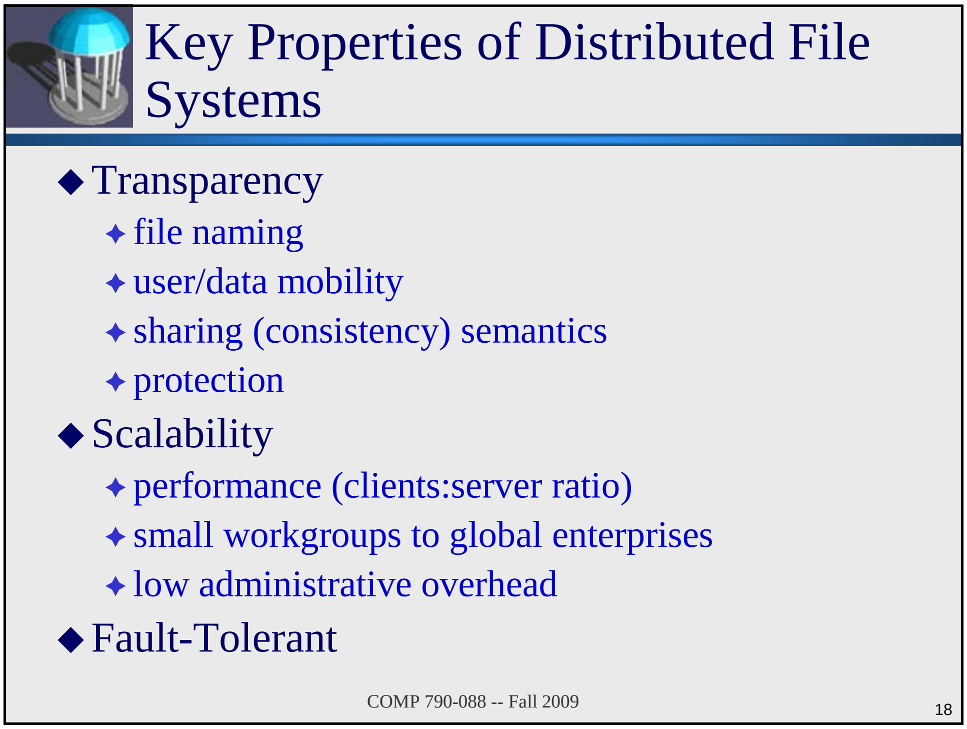
Characterization of File Usage

- File sizes are strongly skewed
  - most files accessed are small
  - most bytes come from large files
- Reads are more frequent than writes (5:1 – 2:1)
- Most files are accesses *sequentially* and/or *entirely*
- Mutation is frequent
  - many file lifetimes are short
  - file data is often modified over short intervals



Characterization of File Usage (continued)

- Sharing modes:
  - file read and written by one user (common)
  - file written by one user, read by many (sometimes)
  - file read and written by multiple users (rare)
- "Working sets" exist
- Characterizations may change with type
  - file *vs* directory
  - system *vs* user



Key Properties of Distributed File Systems

- Transparency
  - file naming
  - user/data mobility
  - sharing (consistency) semantics
  - protection
- Scalability
  - performance (clients:server ratio)
  - small workgroups to global enterprises
  - low administrative overhead
- Fault-Tolerant