

Agni: An Efficient Dual-access File System over Object Storage

Kunal Lillaney, Vasily Tarasov,
David Pease, Randal Burns



JOHNS HOPKINS
UNIVERSITY

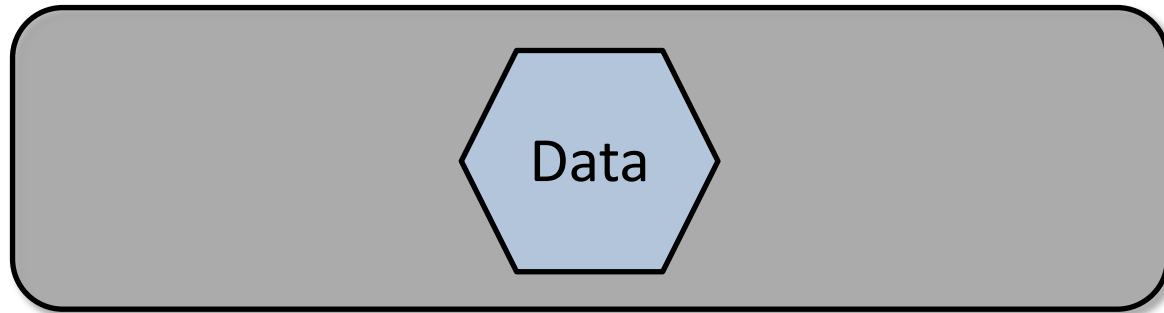
IBM
Research

19 November, IBM Workshop'19 - Almaden

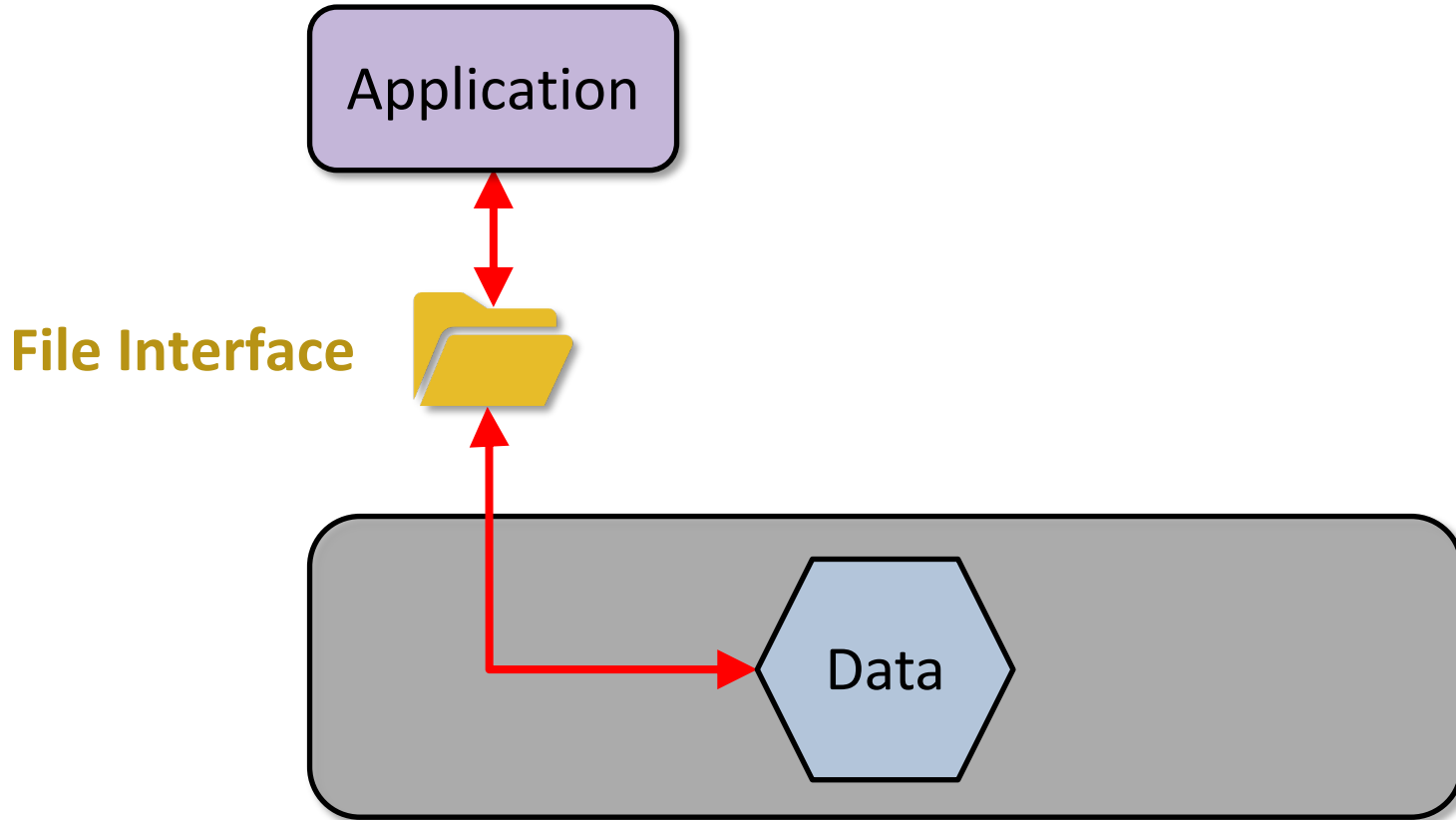
What is Dual-Access?



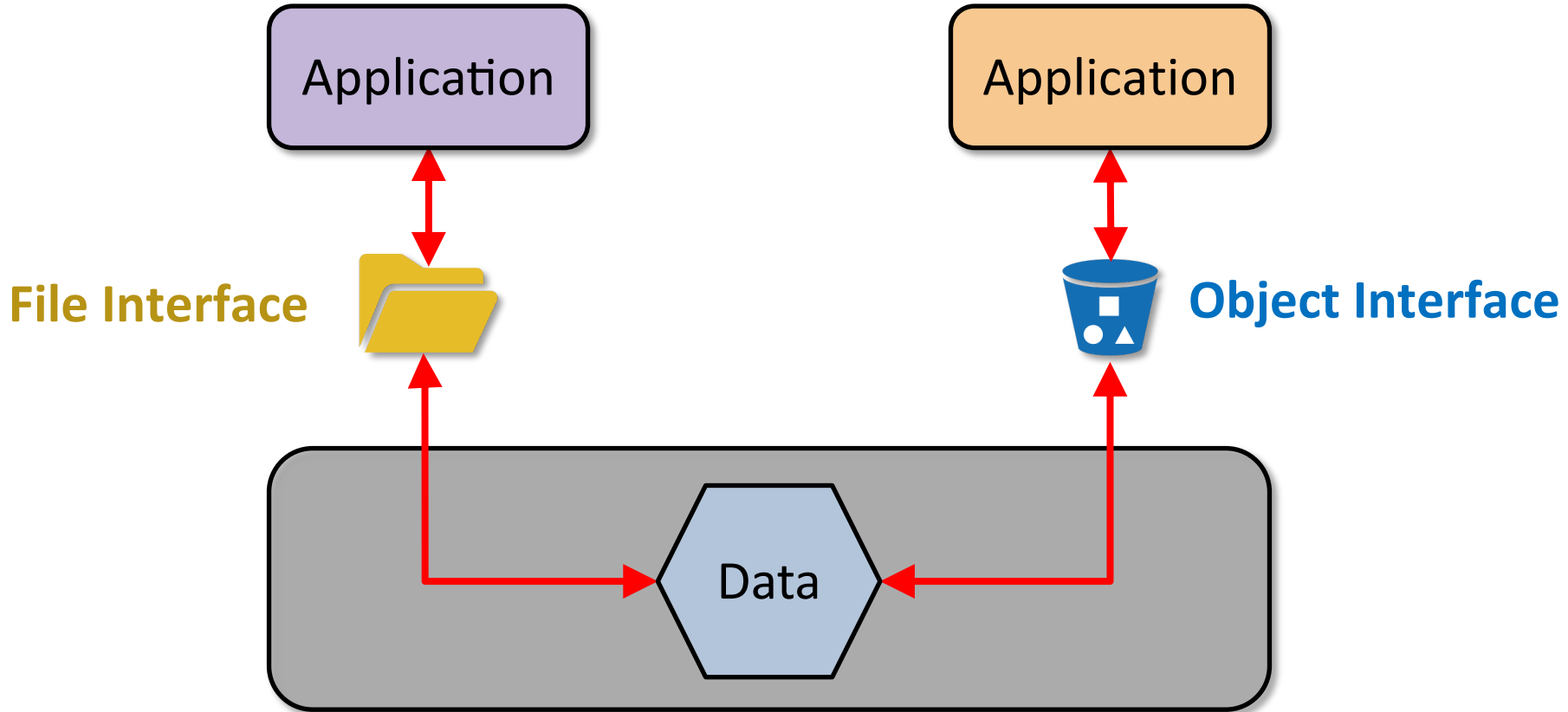
Dual Access



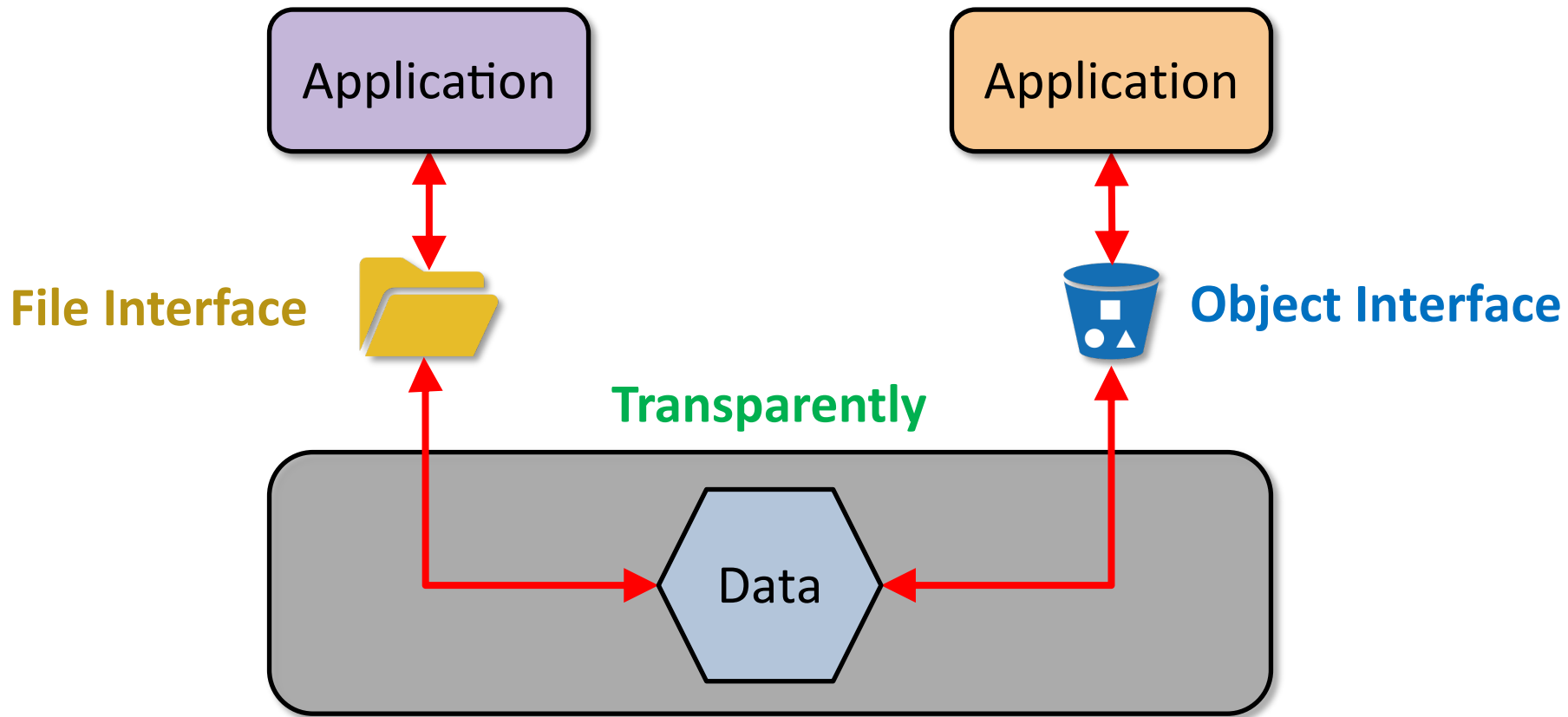
Dual Access



Dual Access



Dual Access





Use
Cases



Media

Use
Cases





Media



Life Science

Use
Cases





Media



Life Science

Use
Cases

Geo-Informatics





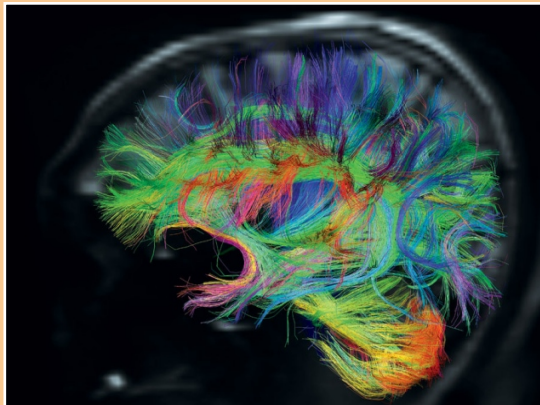
Media



Life Science

Use
Cases

Neuroscience



Geo-Informatics



Media Transcoding, Editing, Analytics



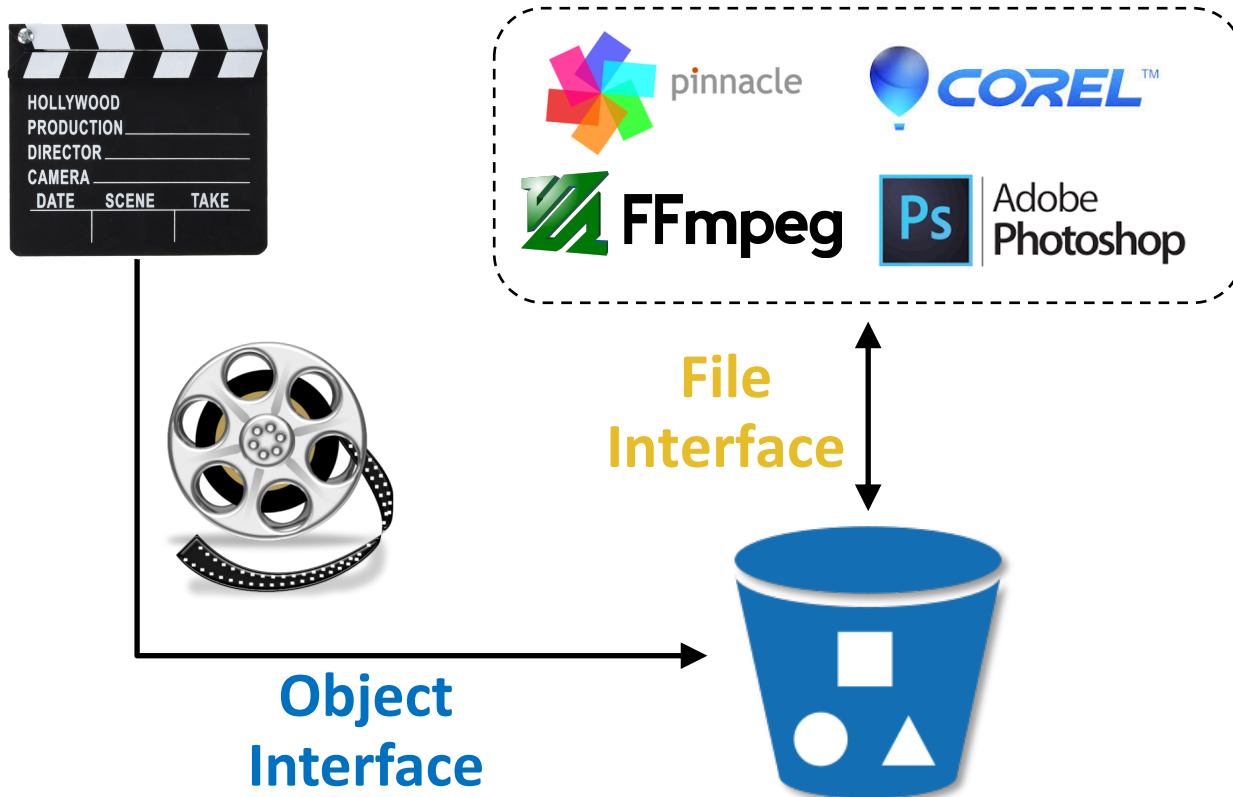
Media Transcoding, Editing, Analytics



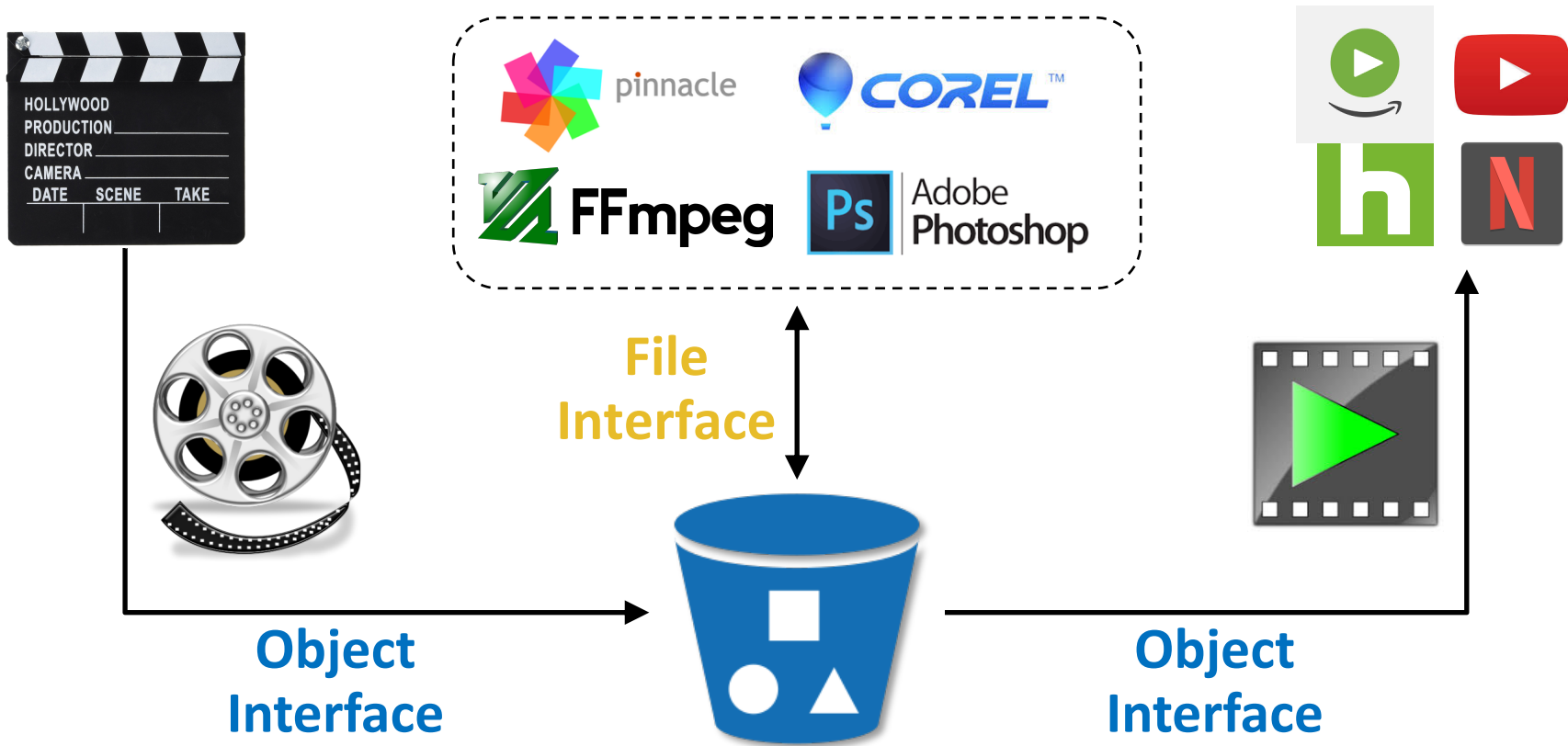
**Object
Interface**



Media Transcoding, Editing, Analytics



Media Transcoding, Editing, Analytics



File Systems vs Object Storage



File Systems vs Object Storage



Partial
Writes



File Systems vs Object Storage



Partial
Writes



File Systems vs Object Storage



Partial
Writes

Namespace

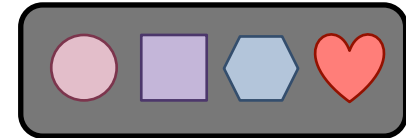
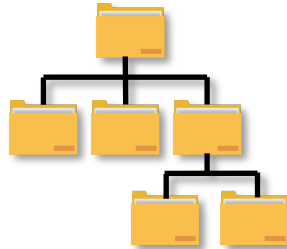


File Systems vs Object Storage



Partial
Writes

Namespace



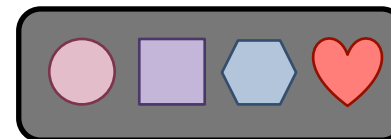
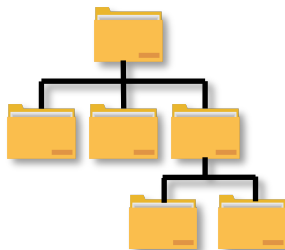
File Systems vs Object Storage



Partial
Writes

Namespace

Interfaces



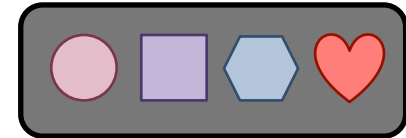
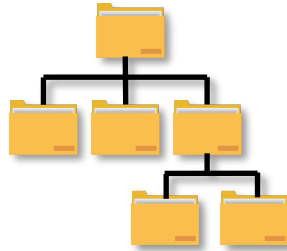
File Systems vs Object Storage



Partial
Writes



Namespace



Interfaces



Outline

- ▶ Design considerations
- ▶ Existing systems
- ▶ Agni
- ▶ Vision



https://www.greenbiz.com/sites/default/files/styles/gbz_article_primary_breakpoints_kalapicture_screenmd_1x/public/images/articles/featured/datacenter_0.jpg?itok=iJm7ezgB×tamp=1483504030



Outline

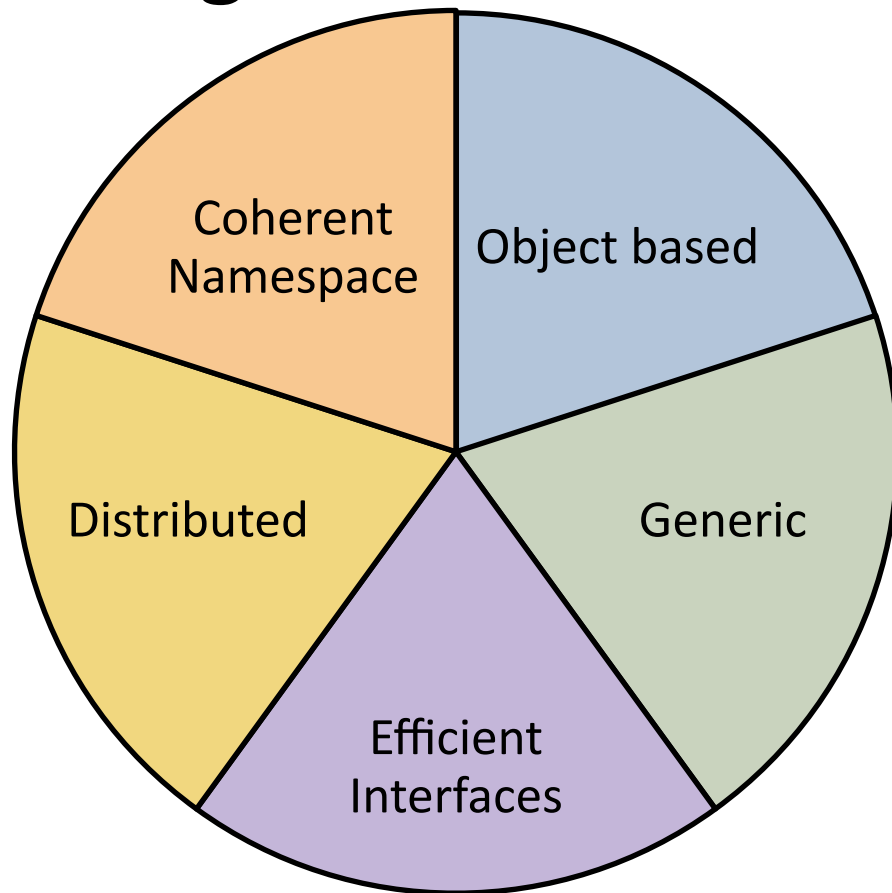
- ▶ Design considerations ←
- ▶ Existing systems
- ▶ Agni
- ▶ Vision



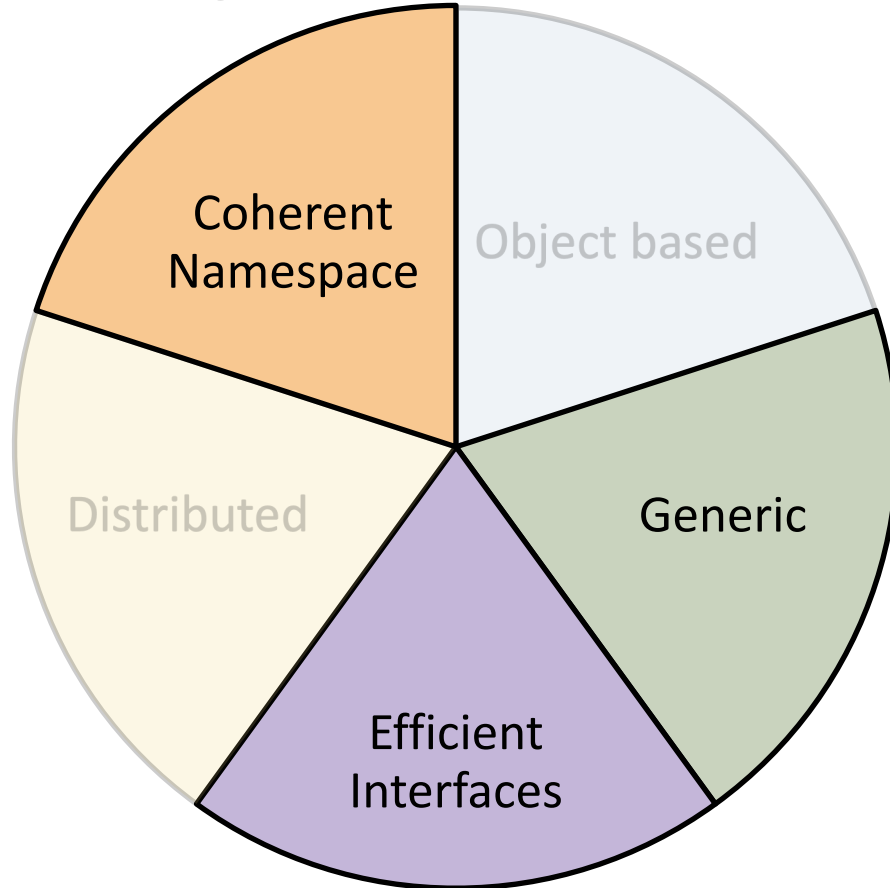
https://www.greenbiz.com/sites/default/files/styles/gbz_article_primary_breakpoints_kalapicture_screenmd_1x/public/images/articles/featured/datacenter_0.jpg?itok=ilm7ezgB×tamp=1483504030



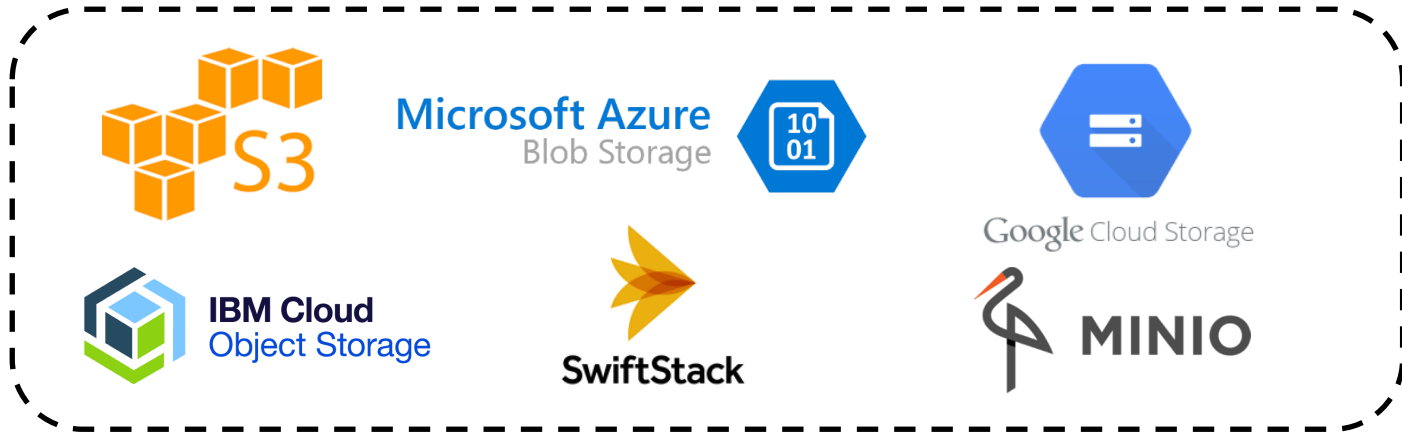
Design Considerations



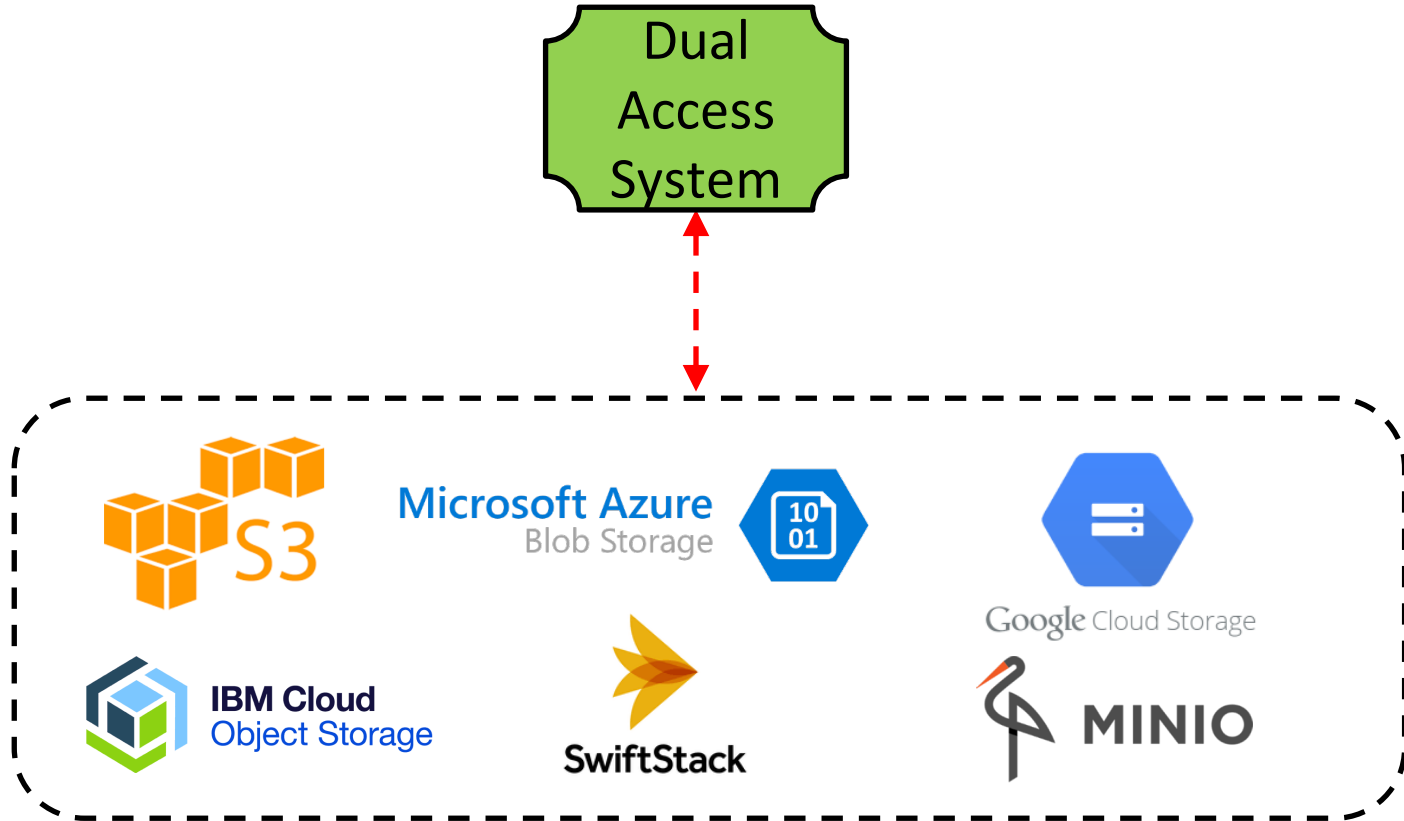
Design Considerations



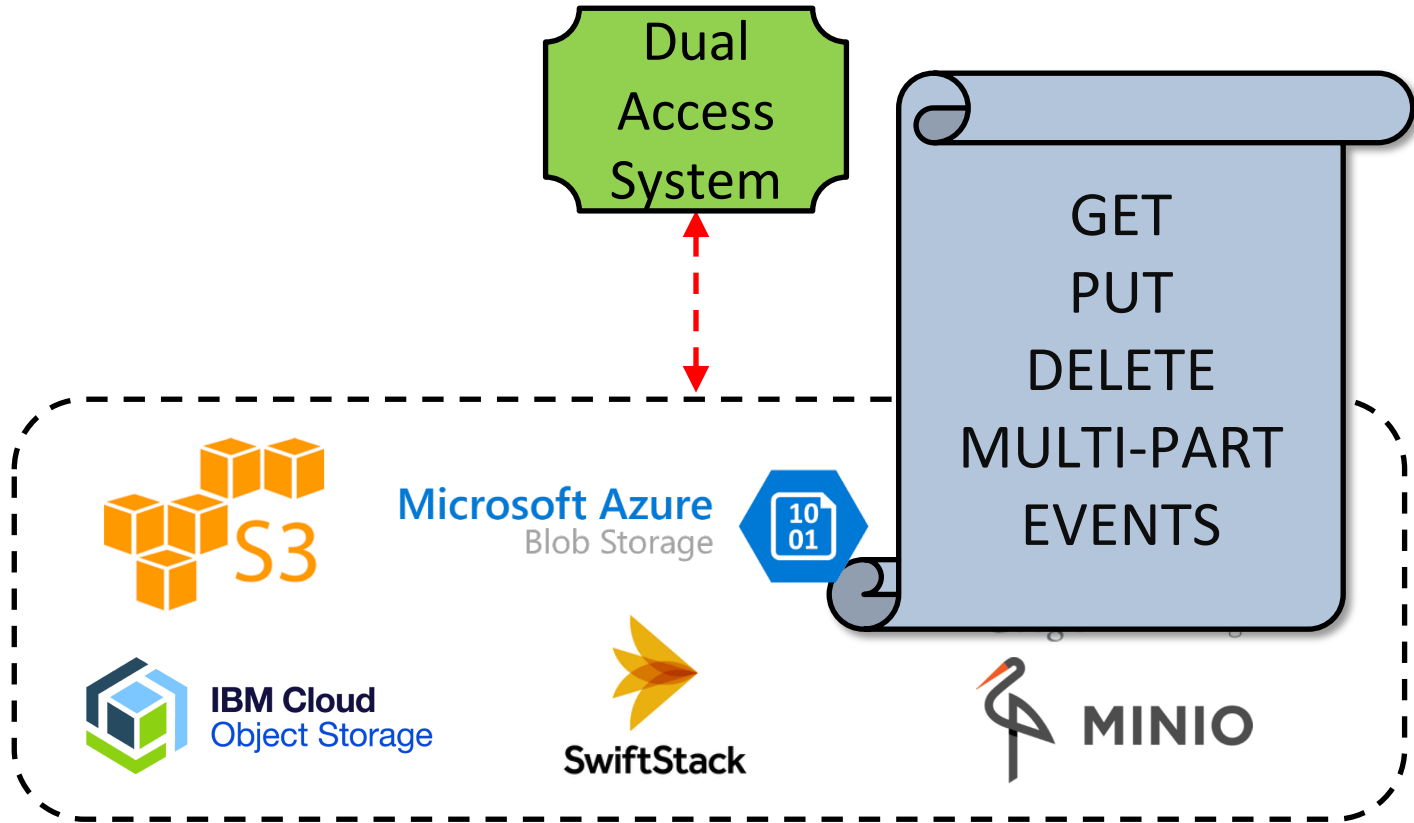
Generic → Object Store Agnostic



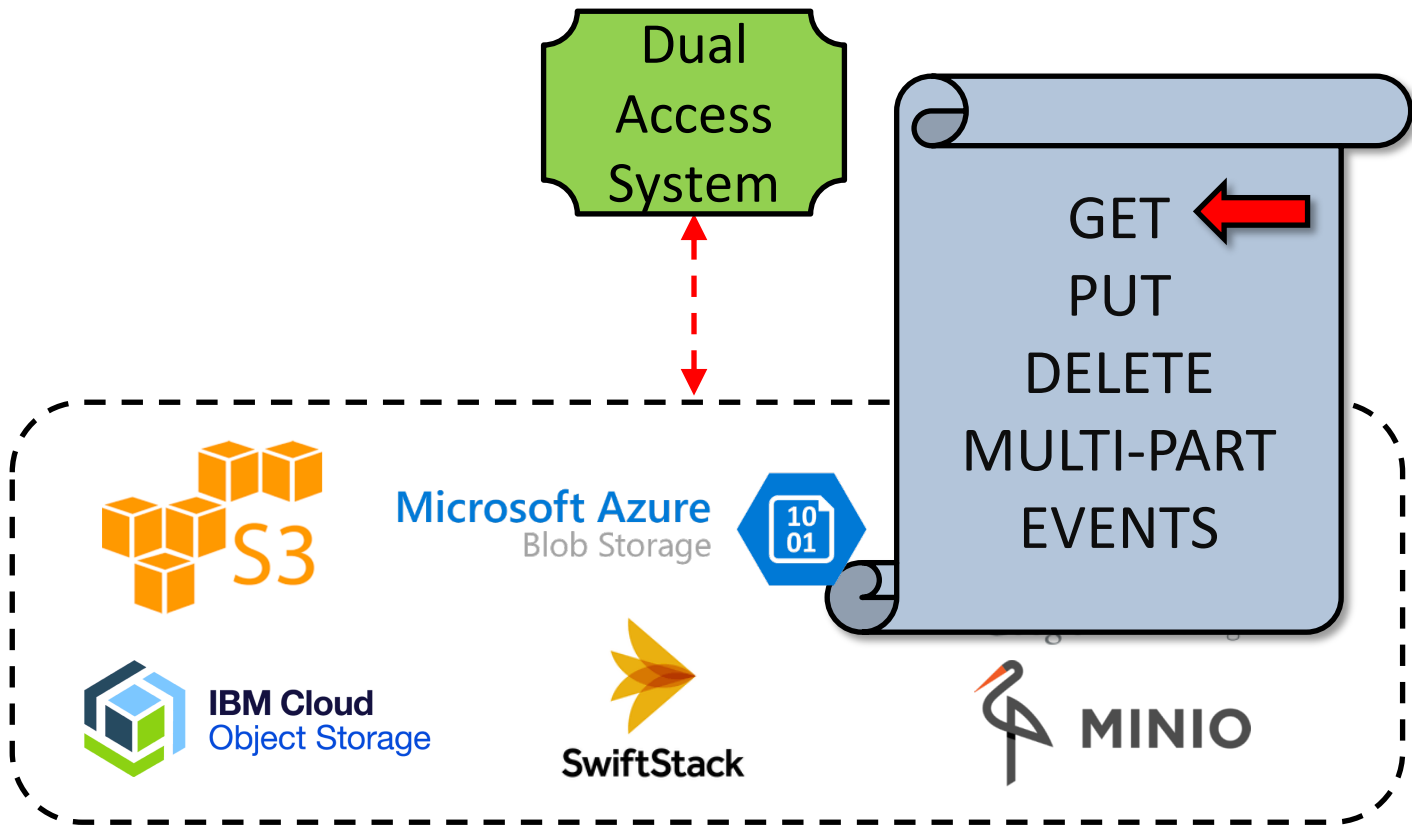
Generic → Object Store Agnostic



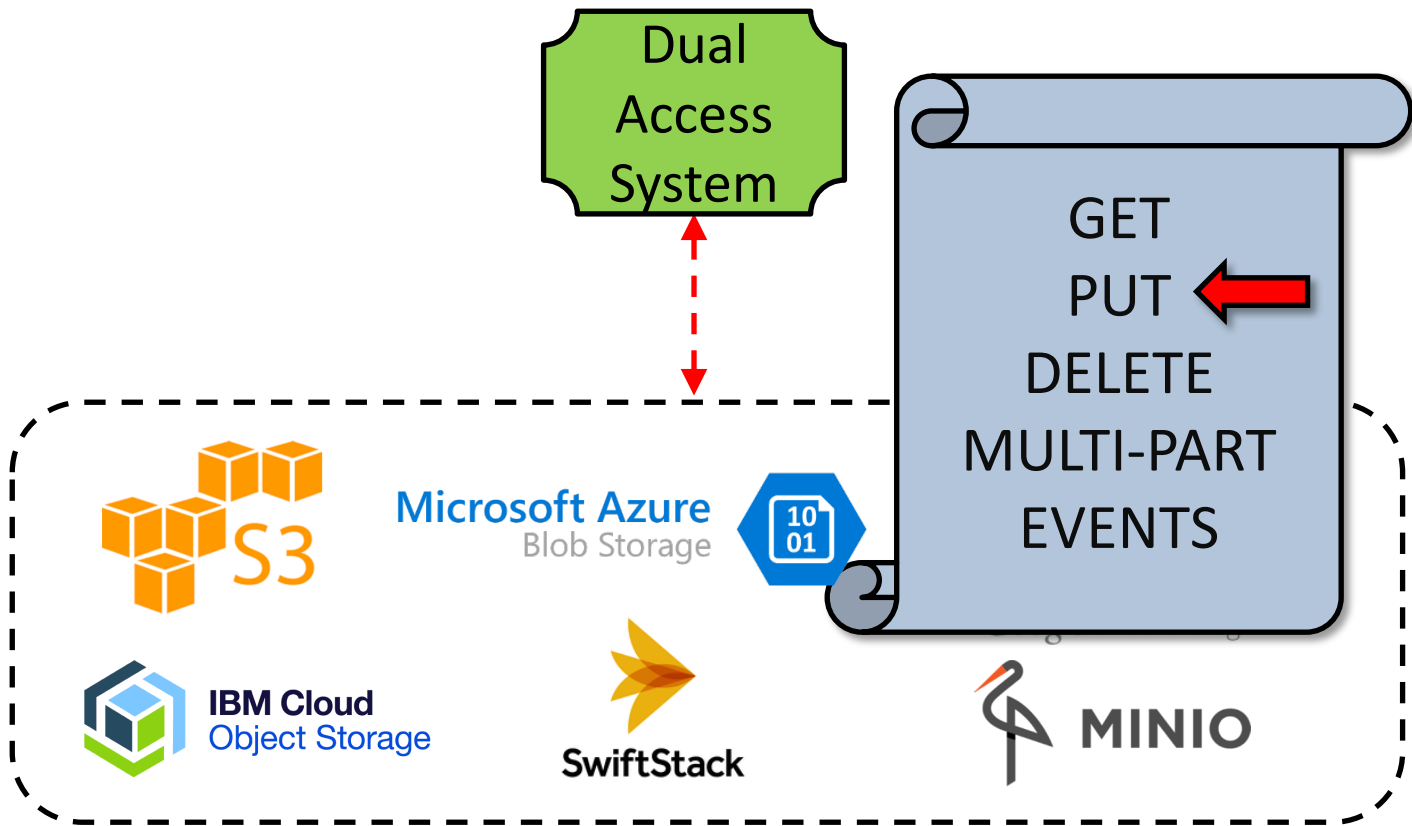
Generic → Object Store Agnostic



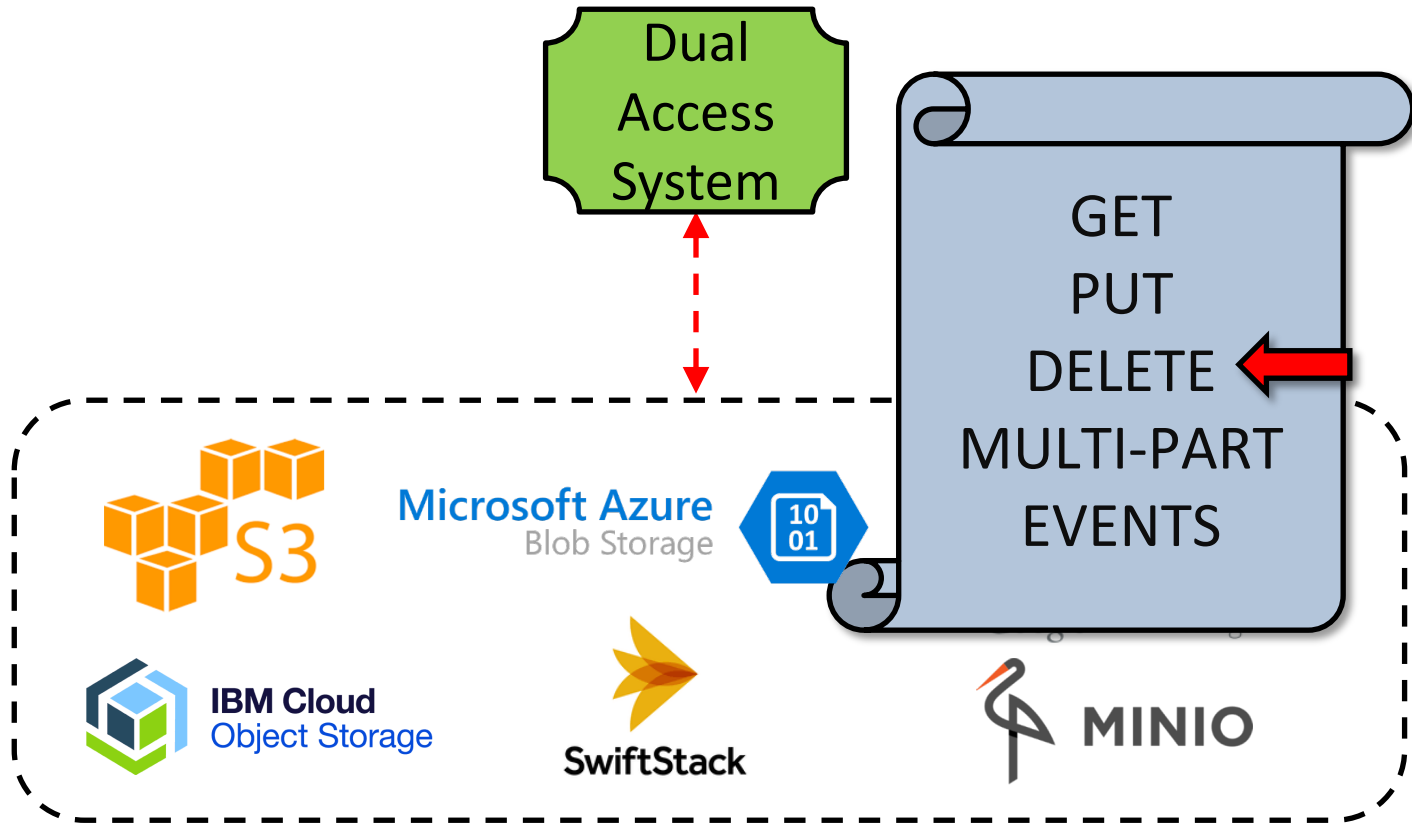
Generic → Object Store Agnostic



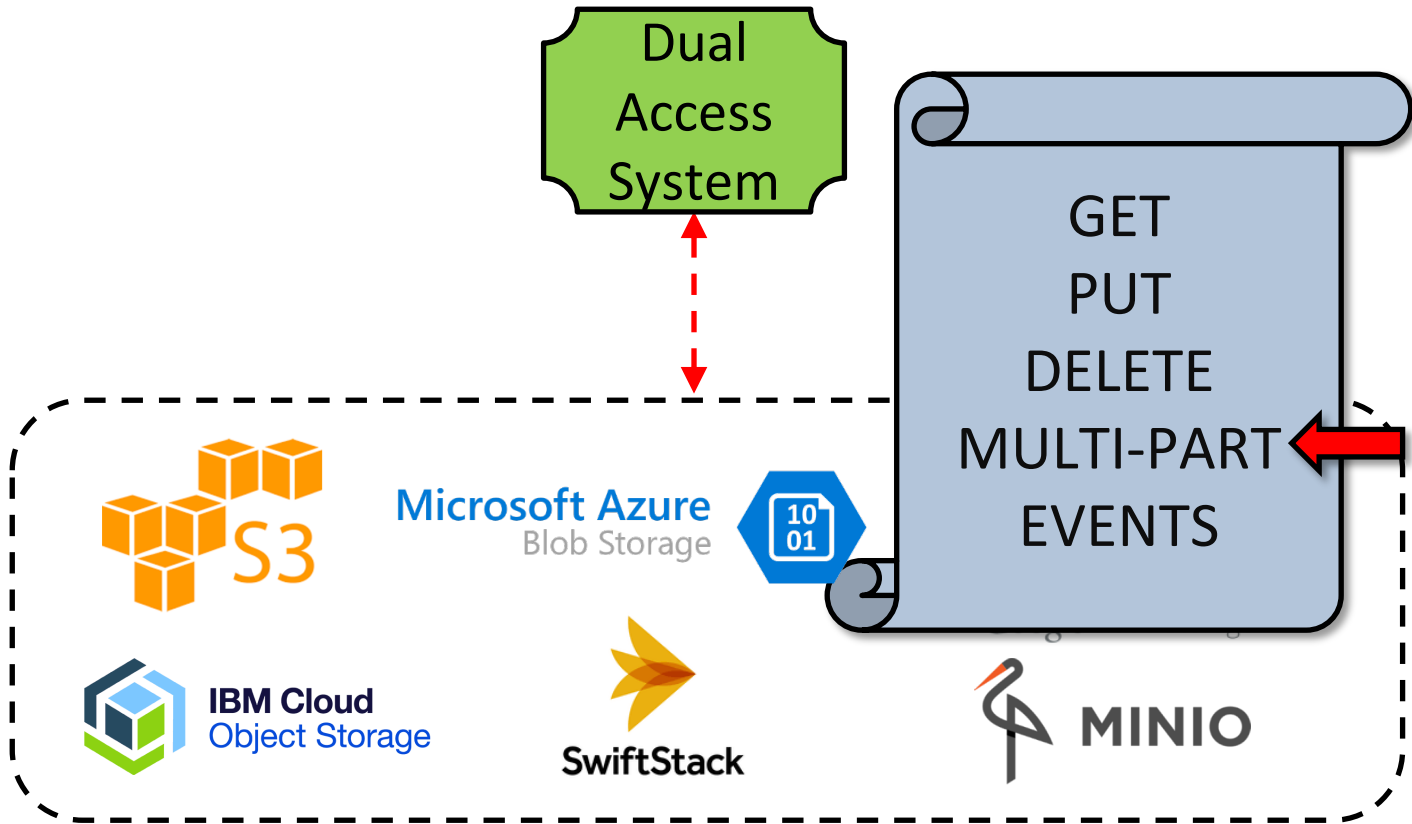
Generic → Object Store Agnostic



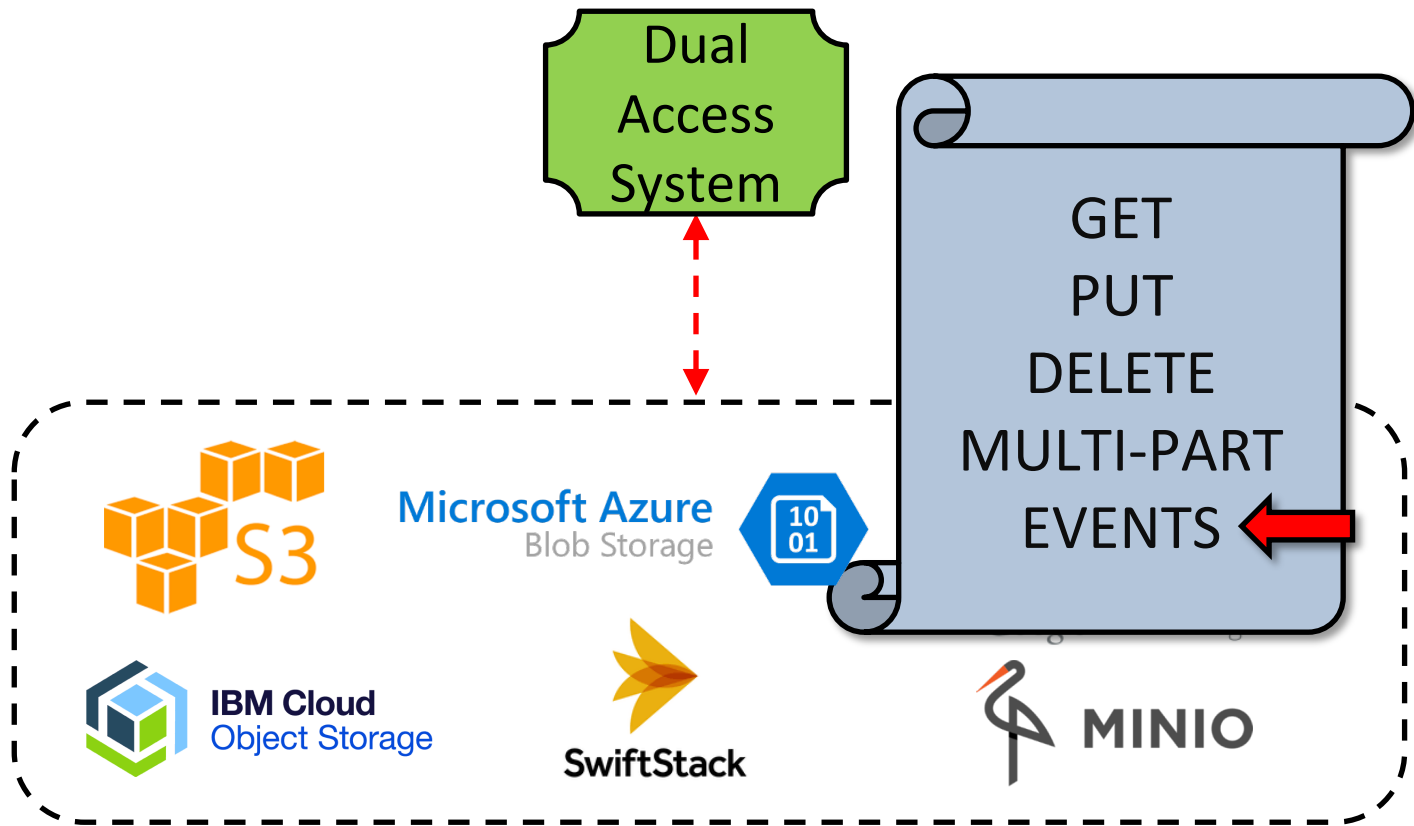
Generic → Object Store Agnostic



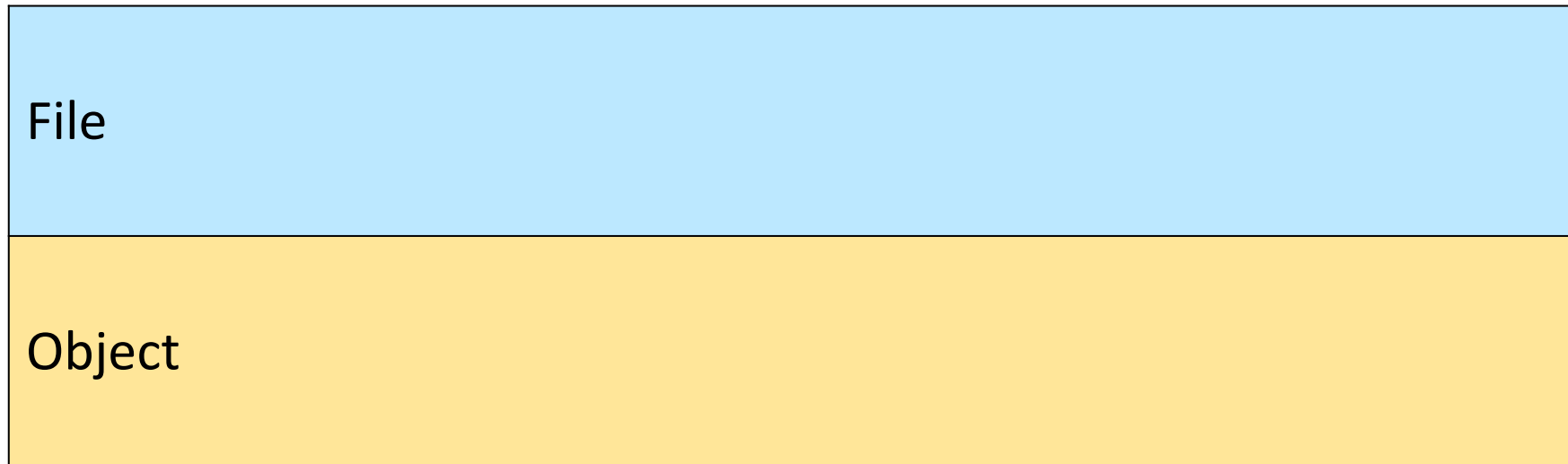
Generic → Object Store Agnostic



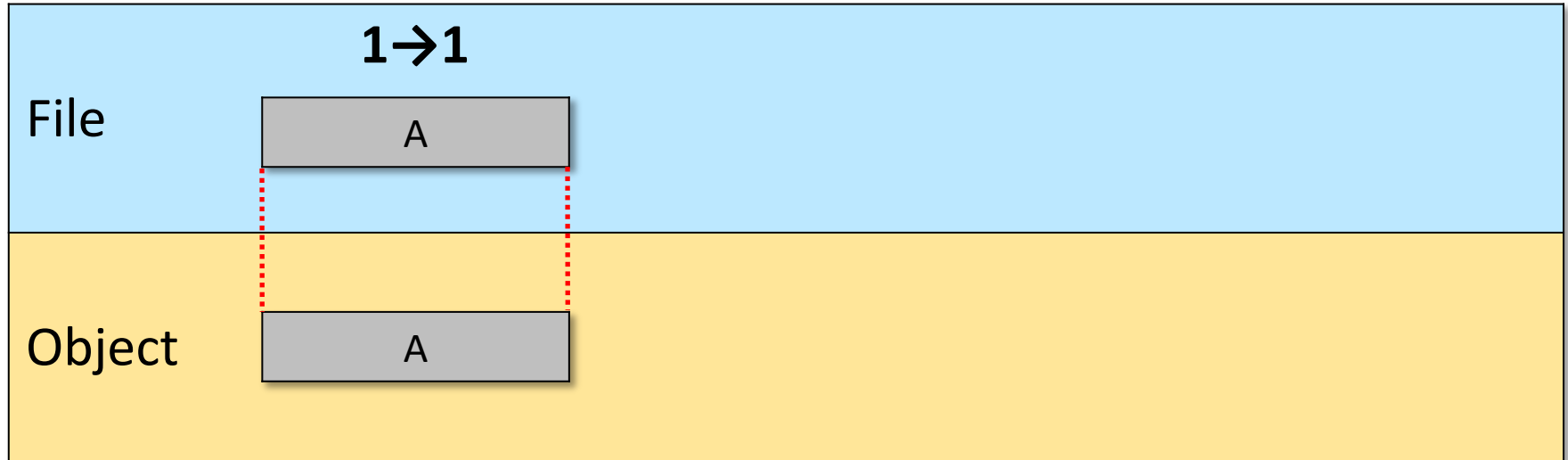
Generic → Object Store Agnostic



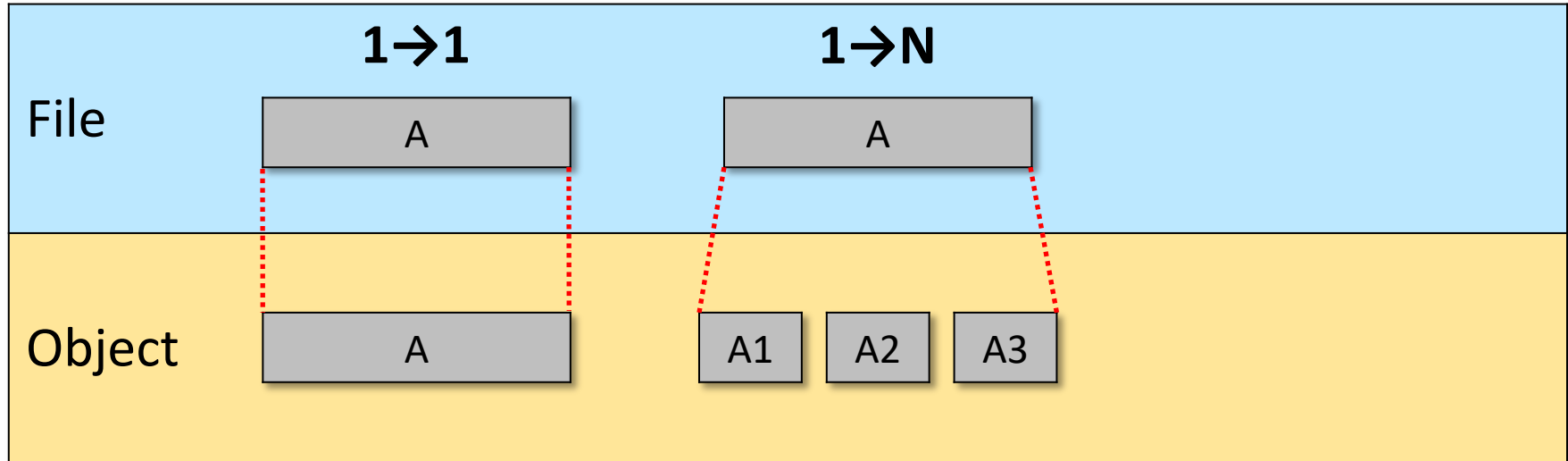
Dual Access: File to Object Mapping



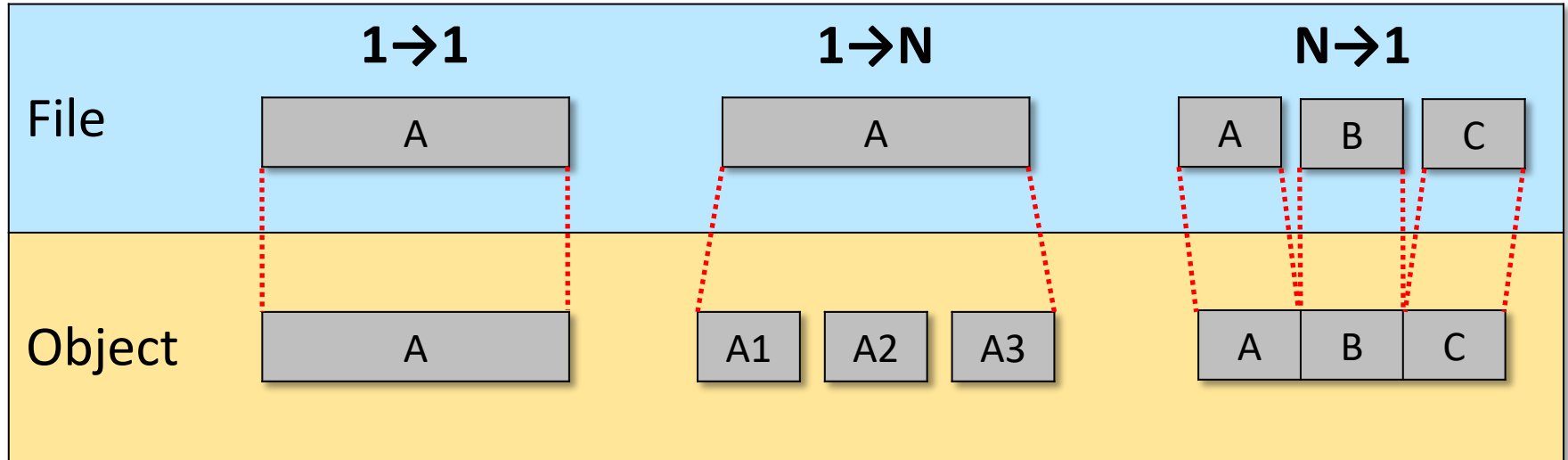
Dual Access: File to Object Mapping



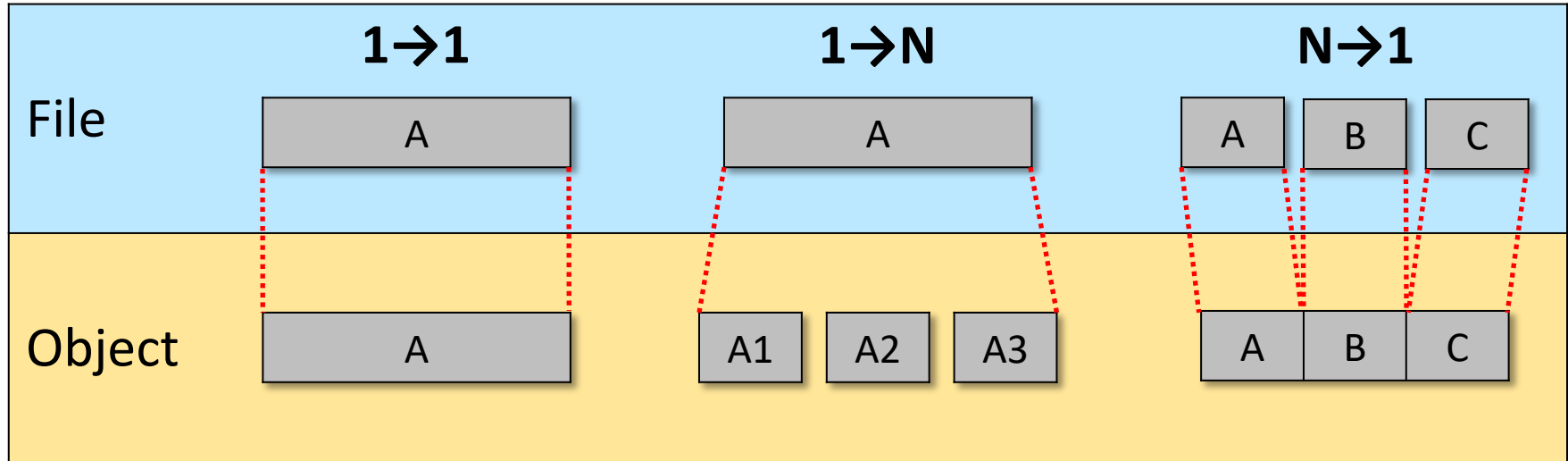
Dual Access: File to Object Mapping



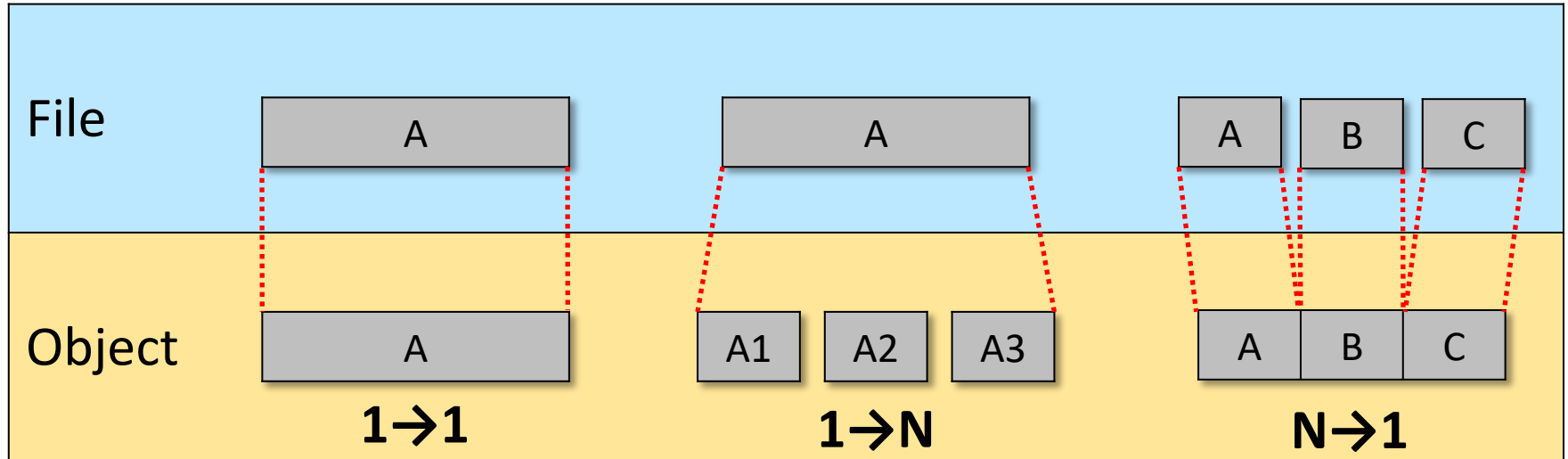
Dual Access: File to Object Mapping



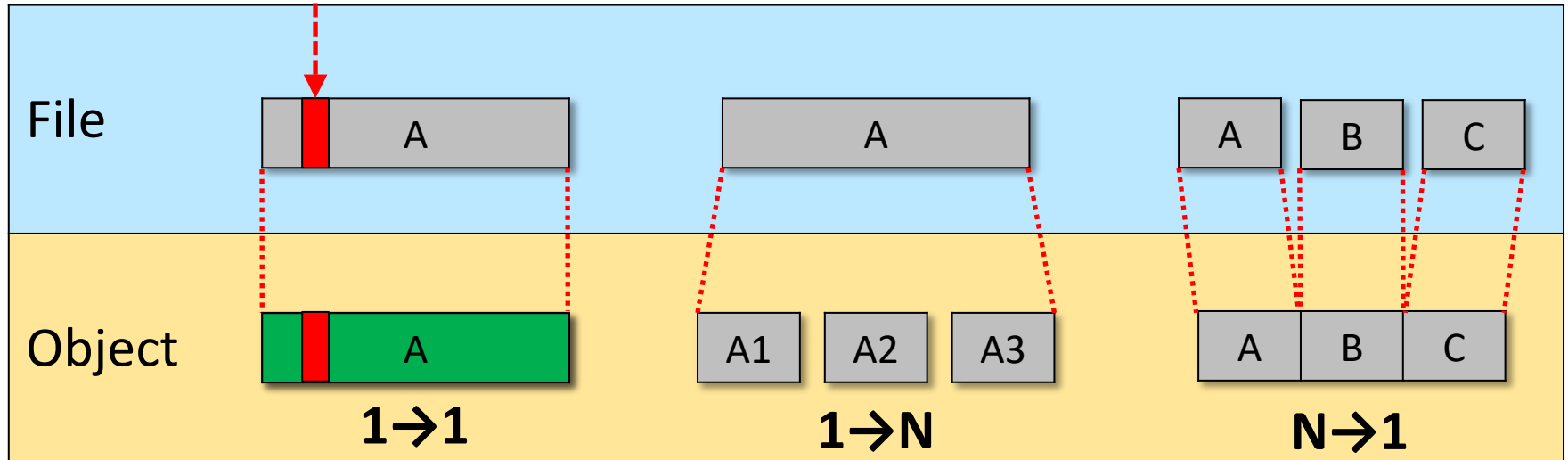
Dual Access: File to Object Mapping



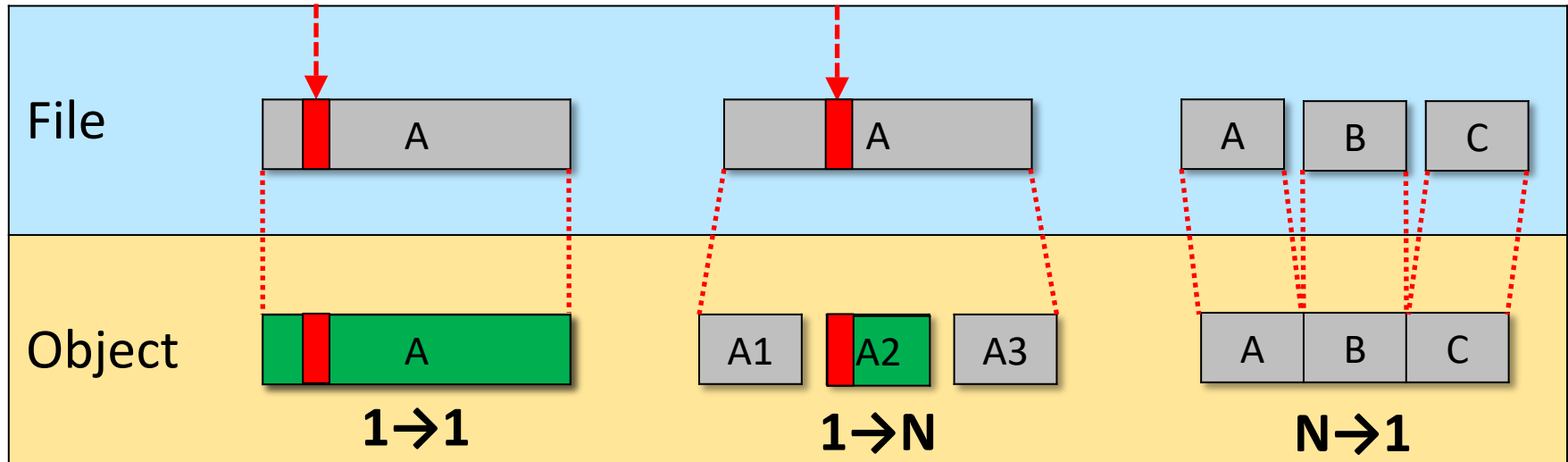
Efficiency: File to Object Mapping



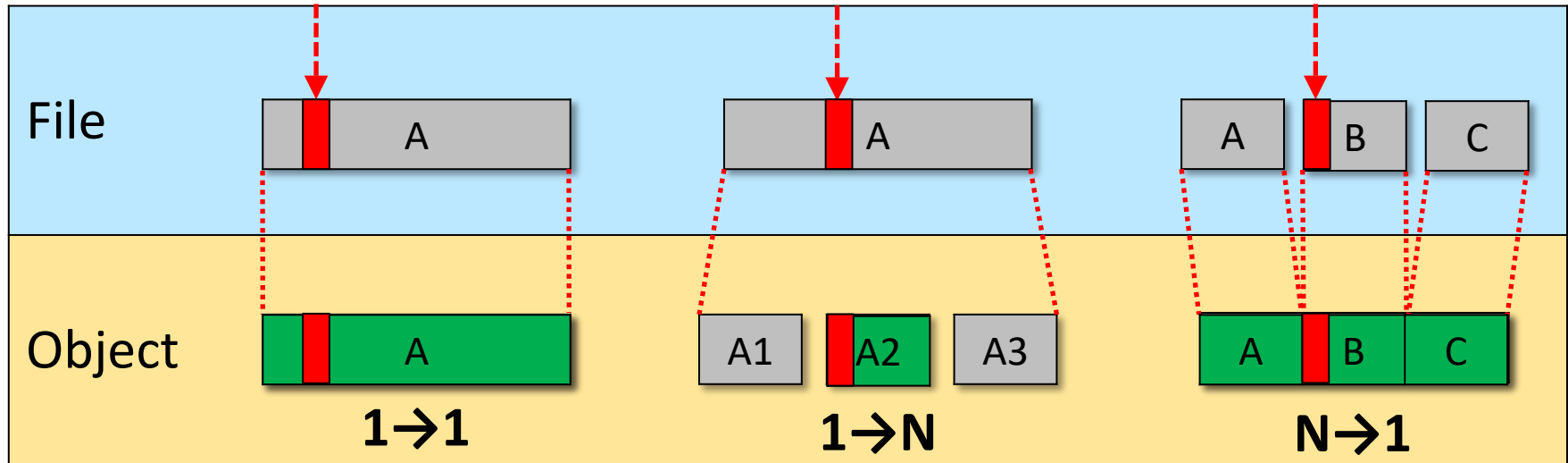
Efficiency: File to Object Mapping



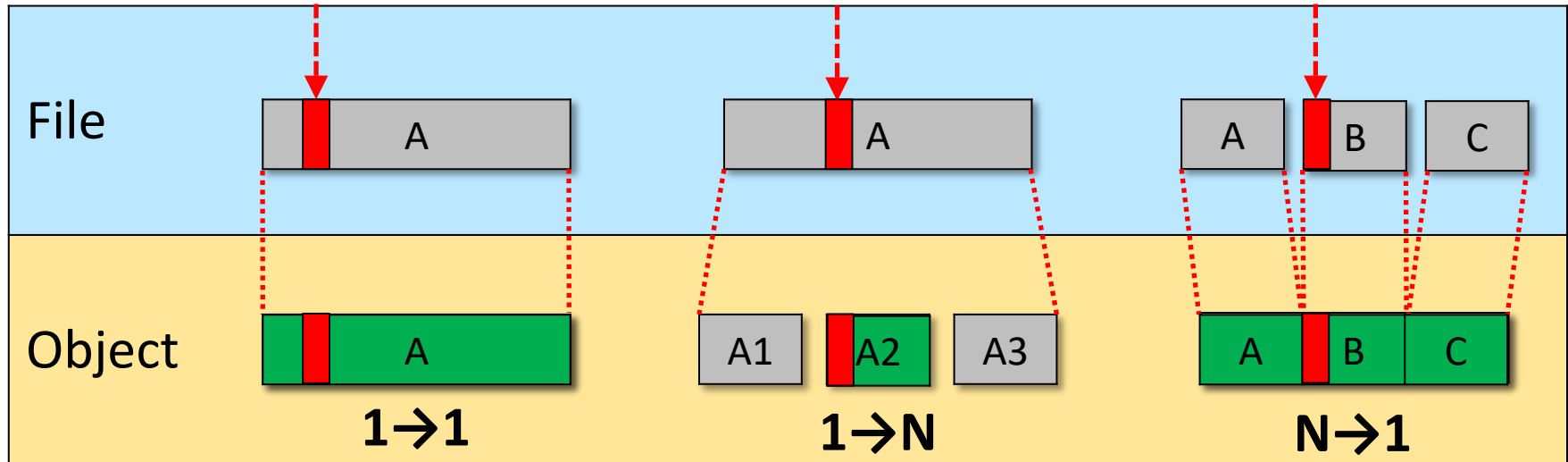
Efficiency: File to Object Mapping



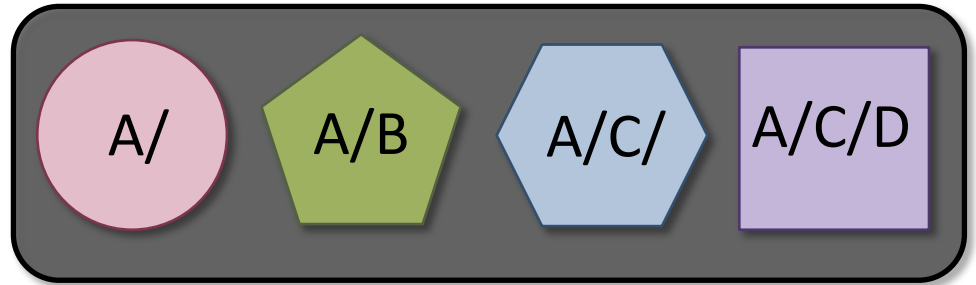
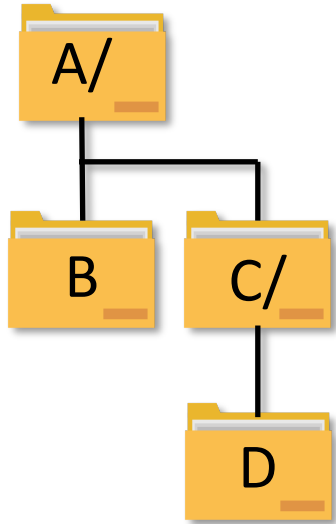
Efficiency: File to Object Mapping



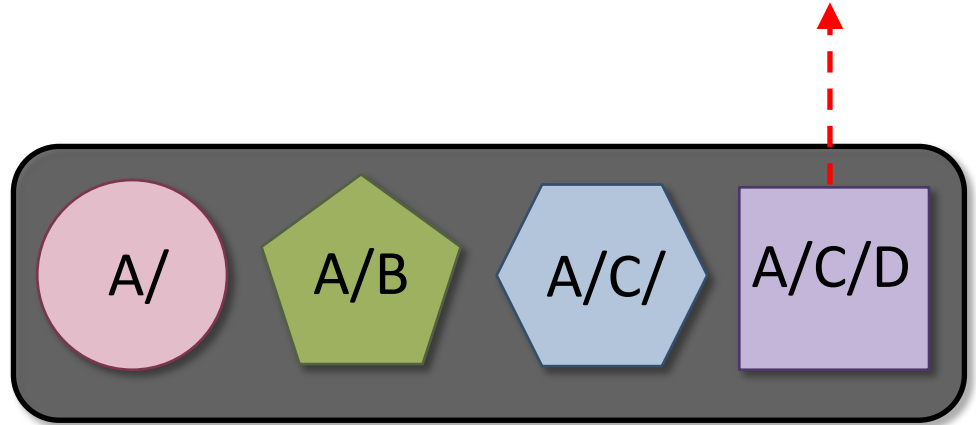
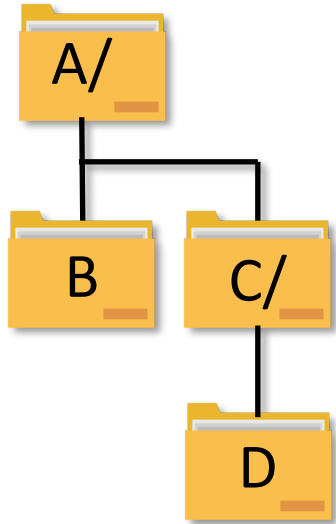
Efficiency: File to Object Mapping



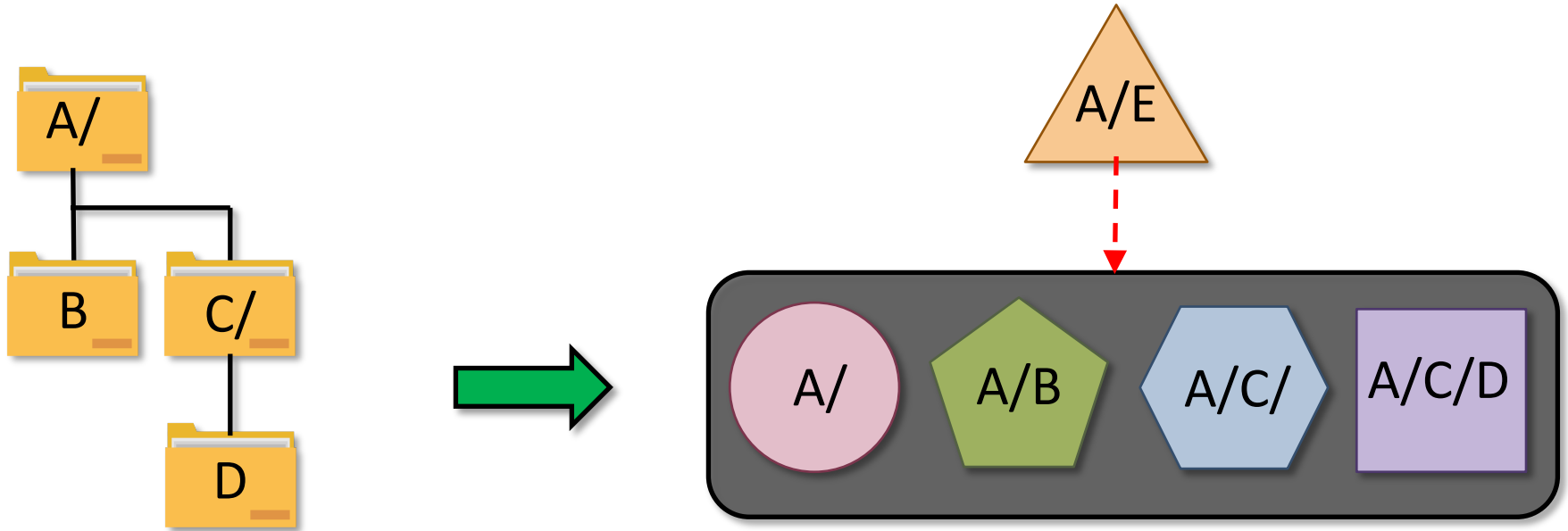
Namespace: Object Naming



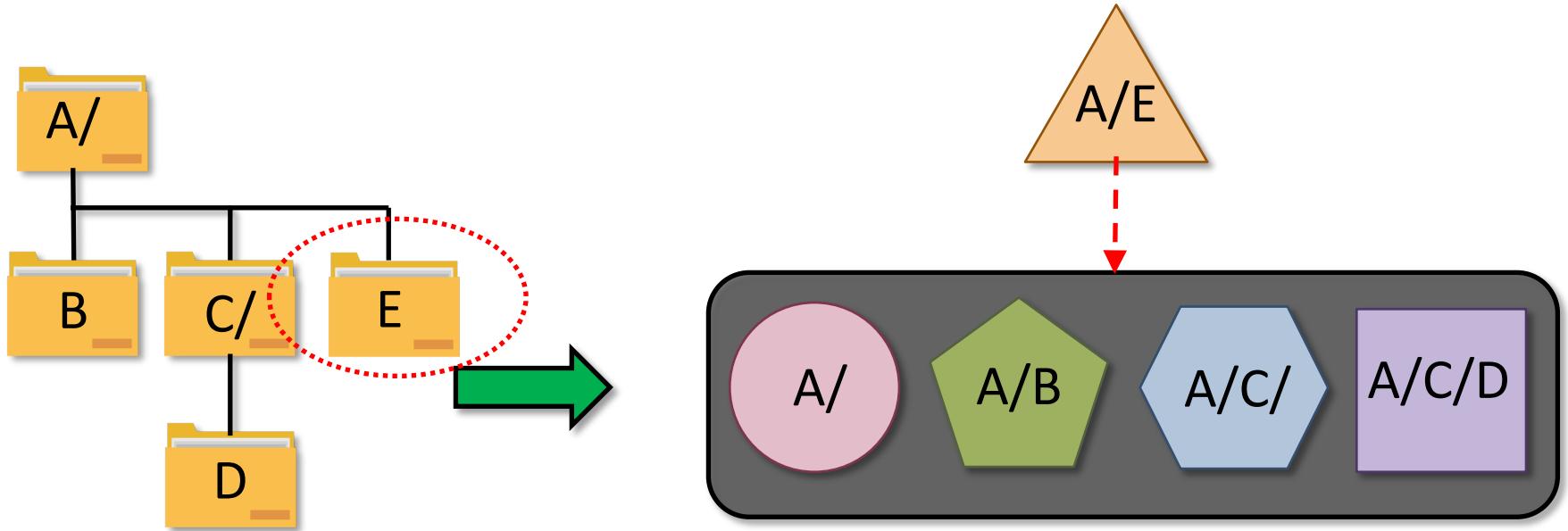
Namespace: Object Naming



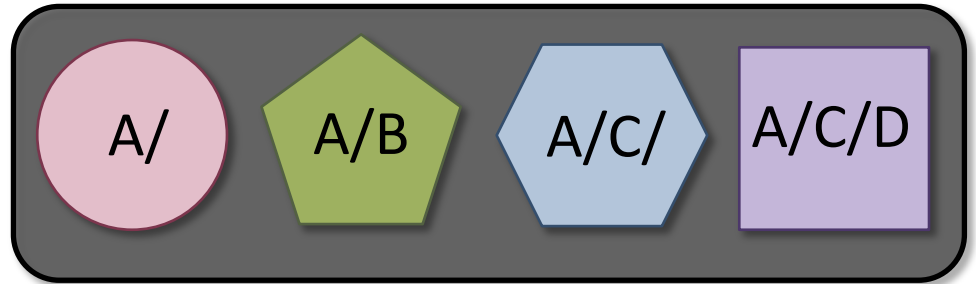
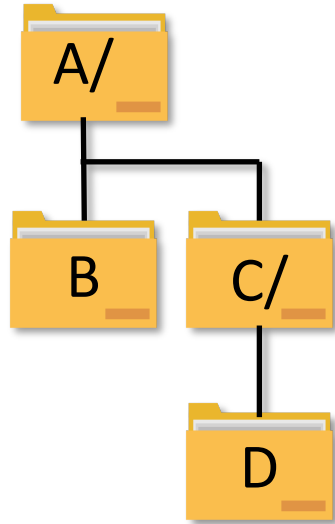
Namespace: Object Naming



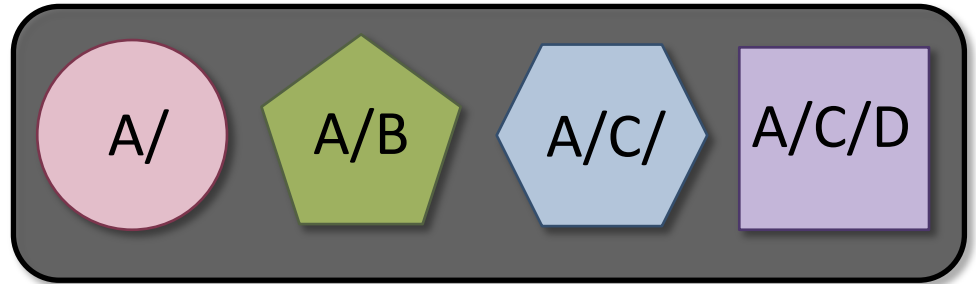
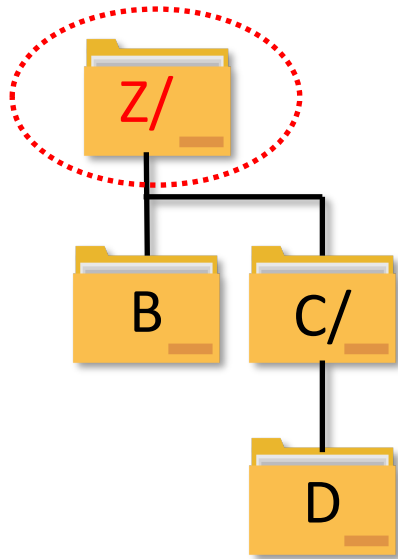
Namespace: Object Naming



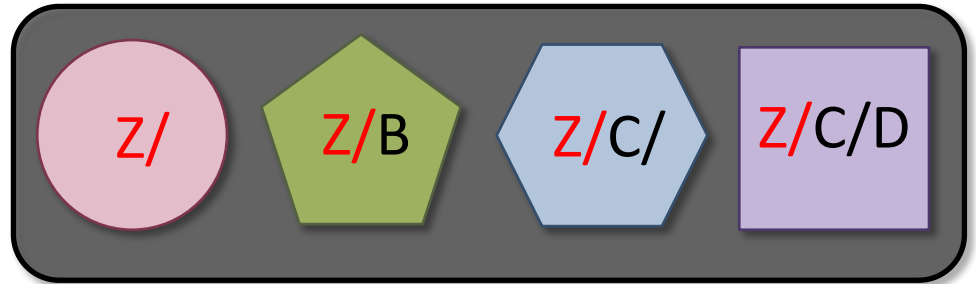
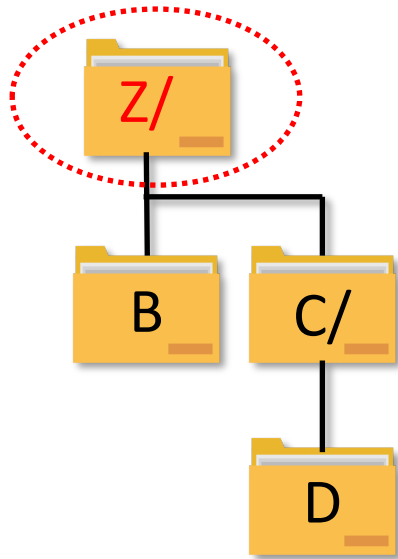
Efficiency: Object Naming



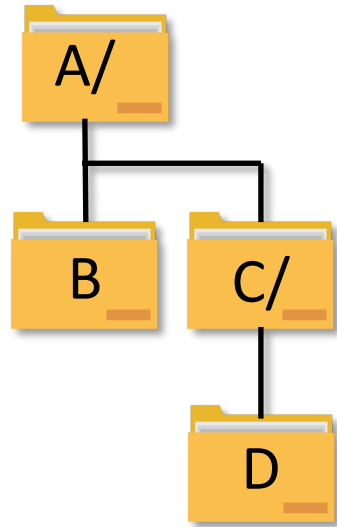
Efficiency: Object Naming



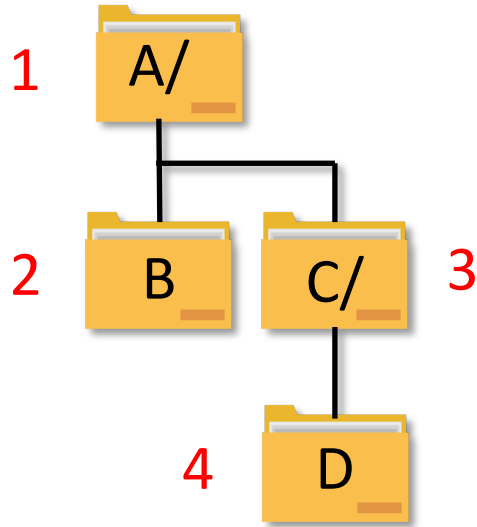
Efficiency: Object Naming



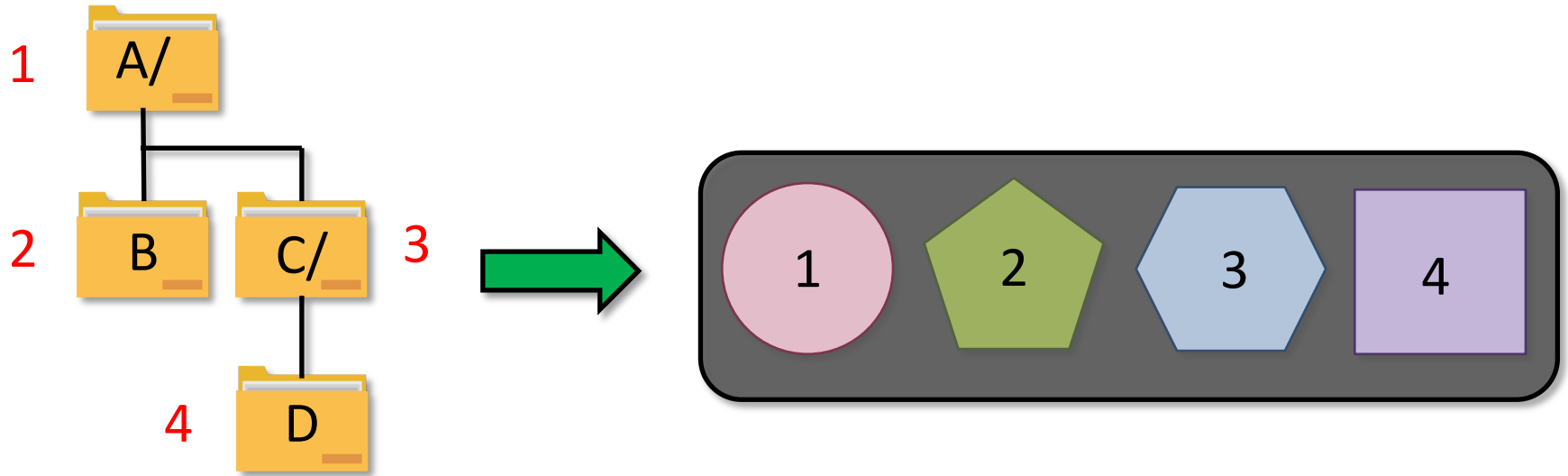
Object Naming: Inode Number



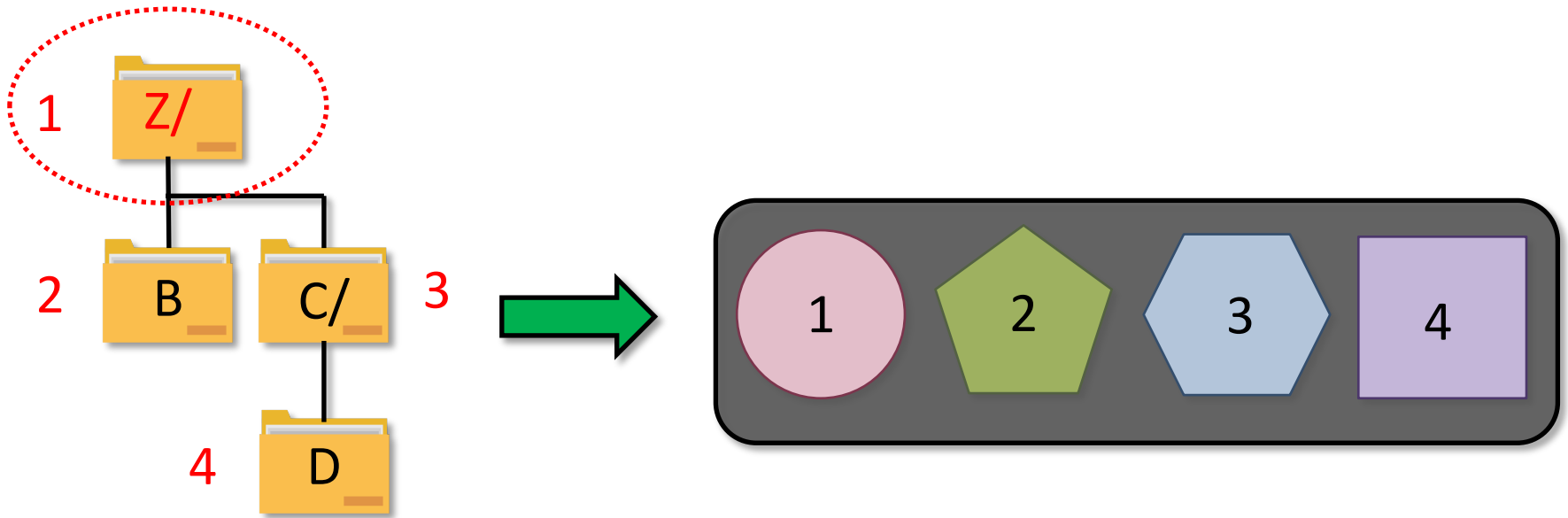
Object Naming: Inode Number



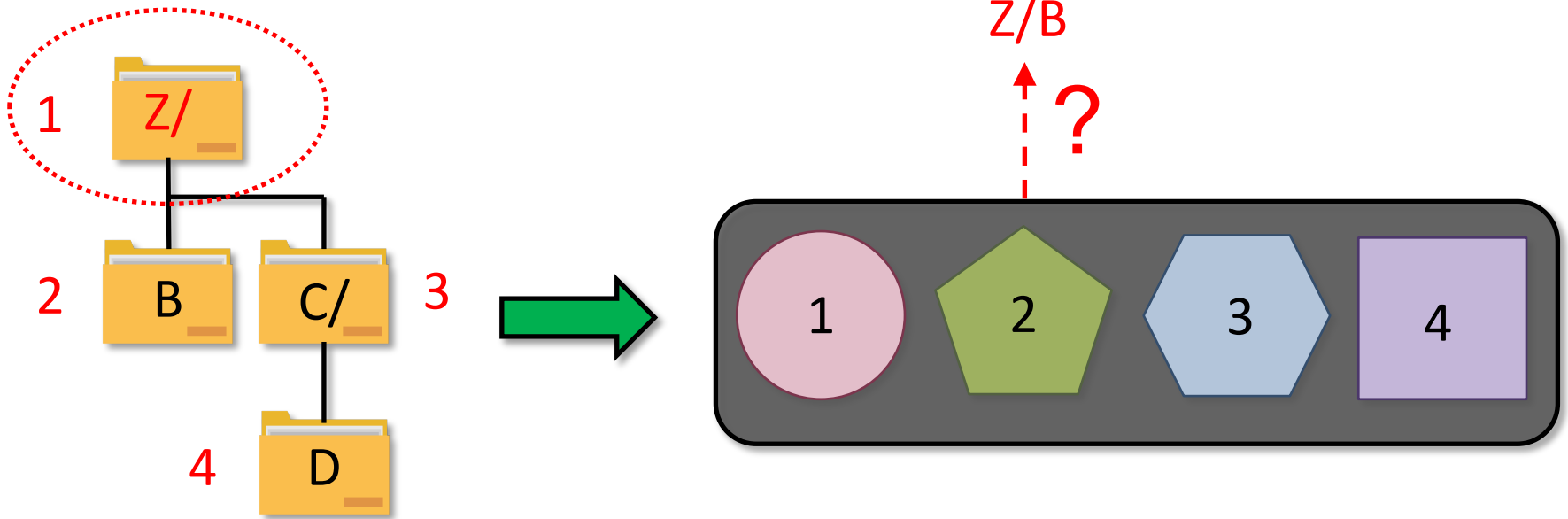
Object Naming: Inode Number



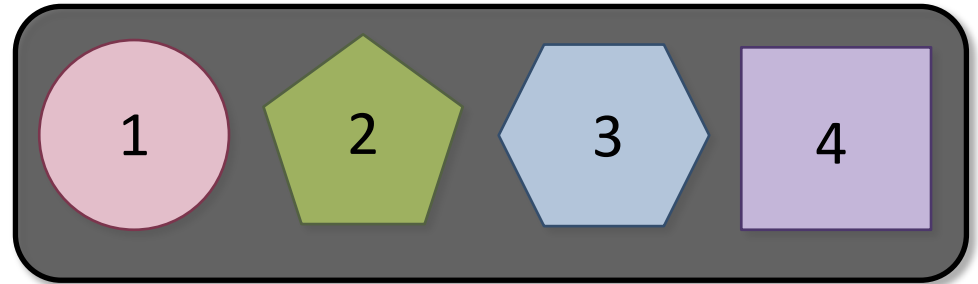
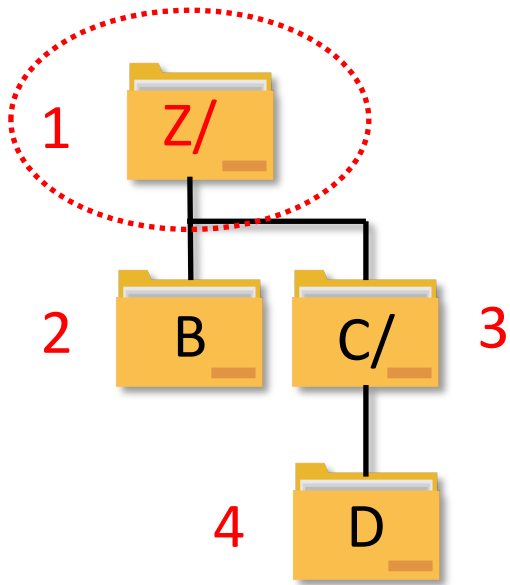
Object Naming: Inode Number



Object Naming: Inode Number

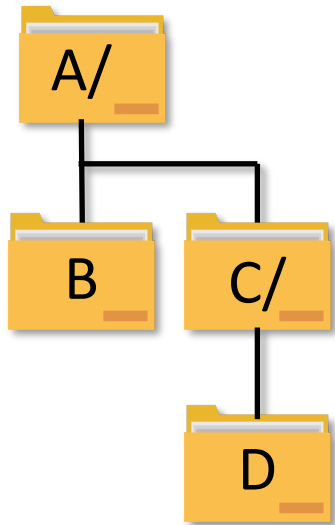


Object Naming: Inode Number

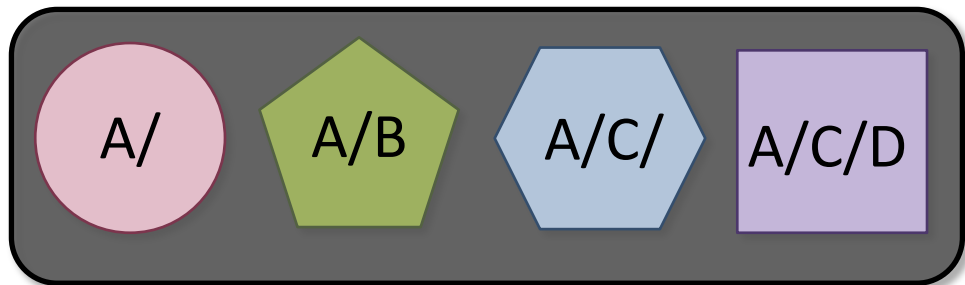
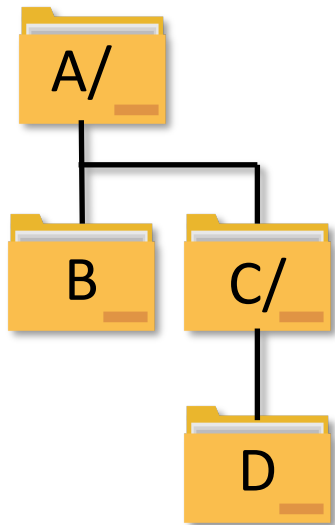


Name	Inode
Z/B	2
...	...

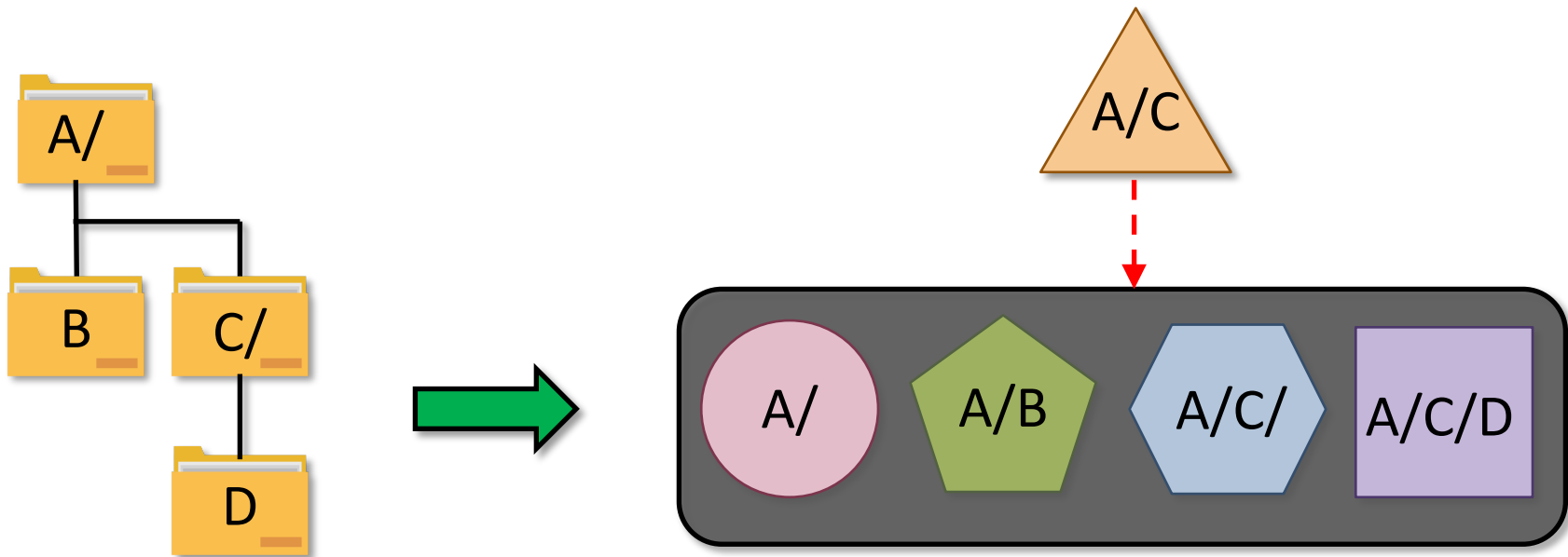
Namespace: Example of Incoherency



Namespace: Example of Incoherency



Namespace: Example of Incoherency



Outline

- ▶ Design considerations
- ▶ Existing systems ←
- ▶ Agni
- ▶ Future work



https://www.greenbiz.com/sites/default/files/styles/gbz_article_primary_breakpoints_kalapicture_screenmd_1x/public/images/articles/featured/datacenter_0.jpg?itok=iJm7ezgB×tamp=1483504030



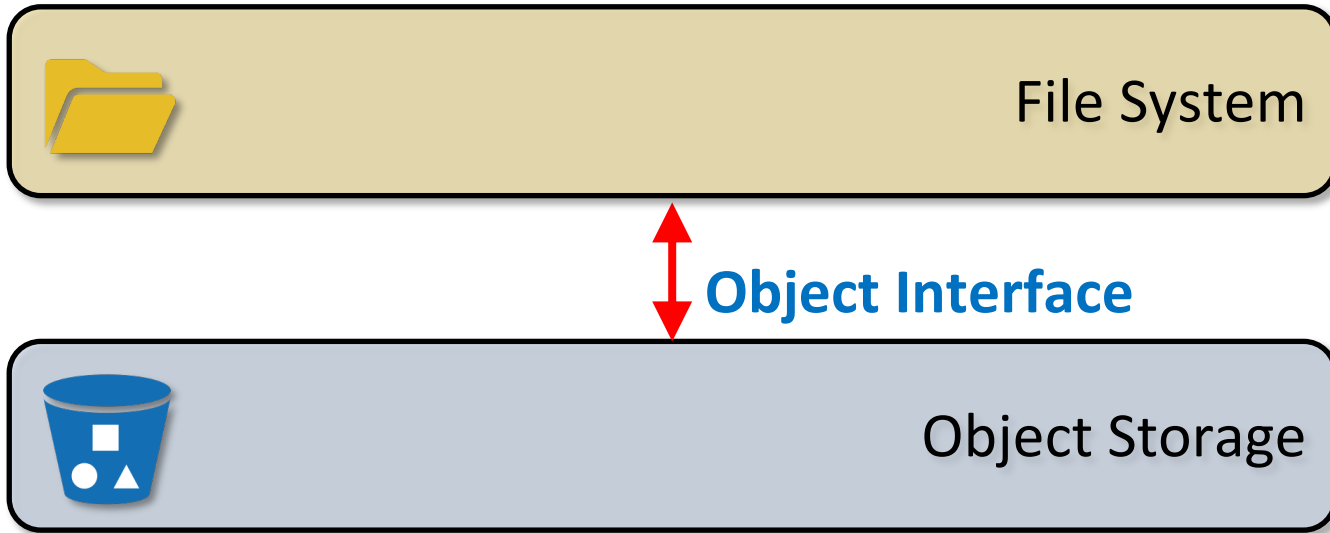
File Systems Paired With Object Storage



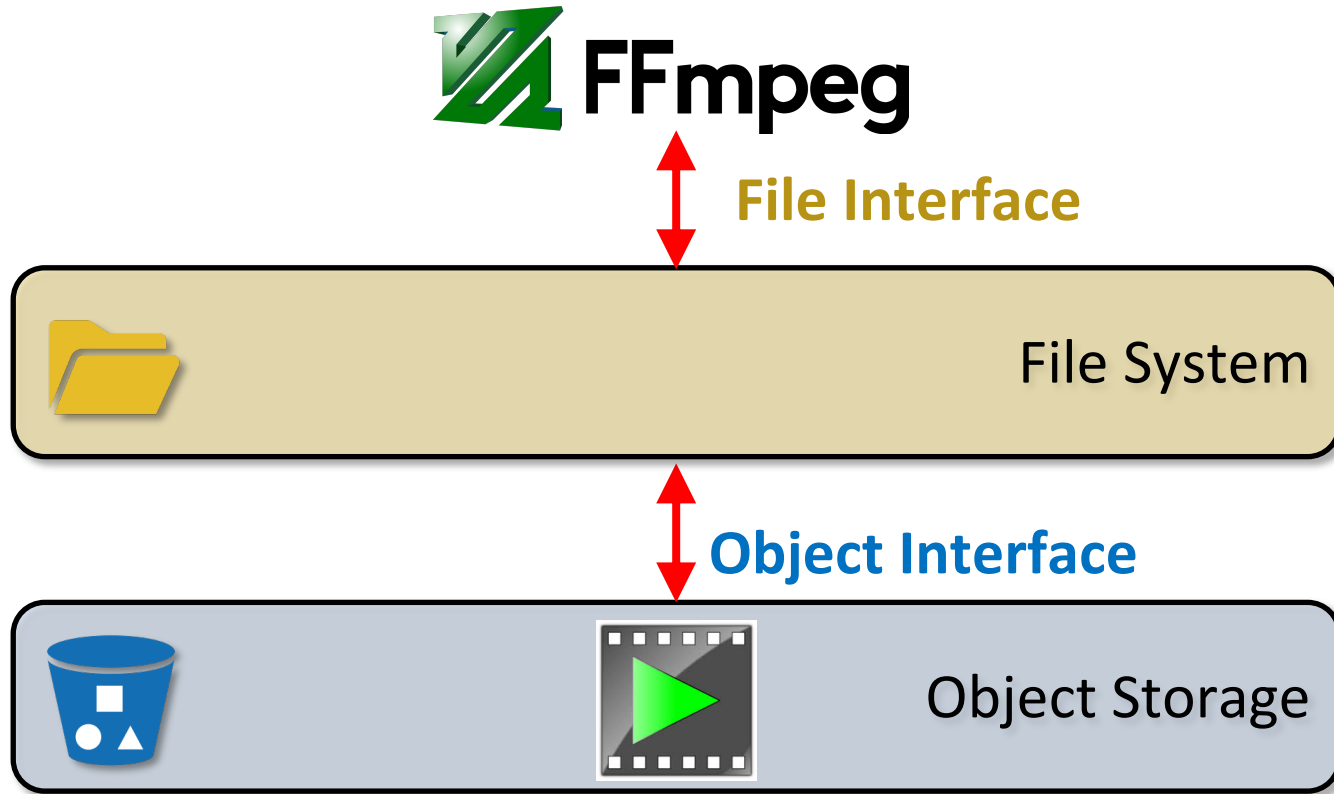
 Object Interface



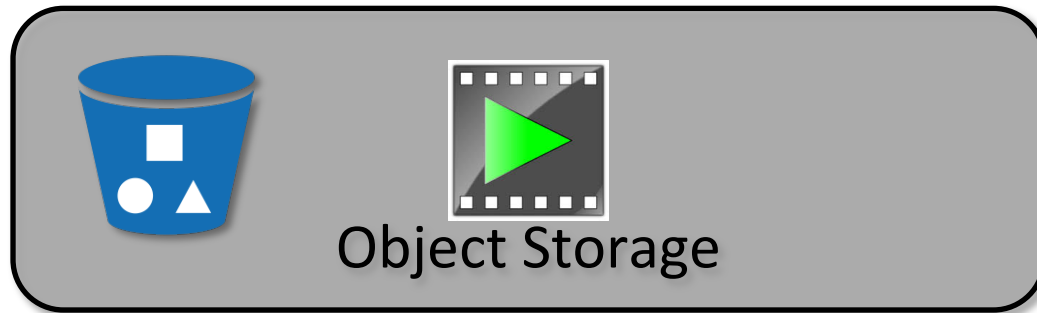
File Systems Paired With Object Storage



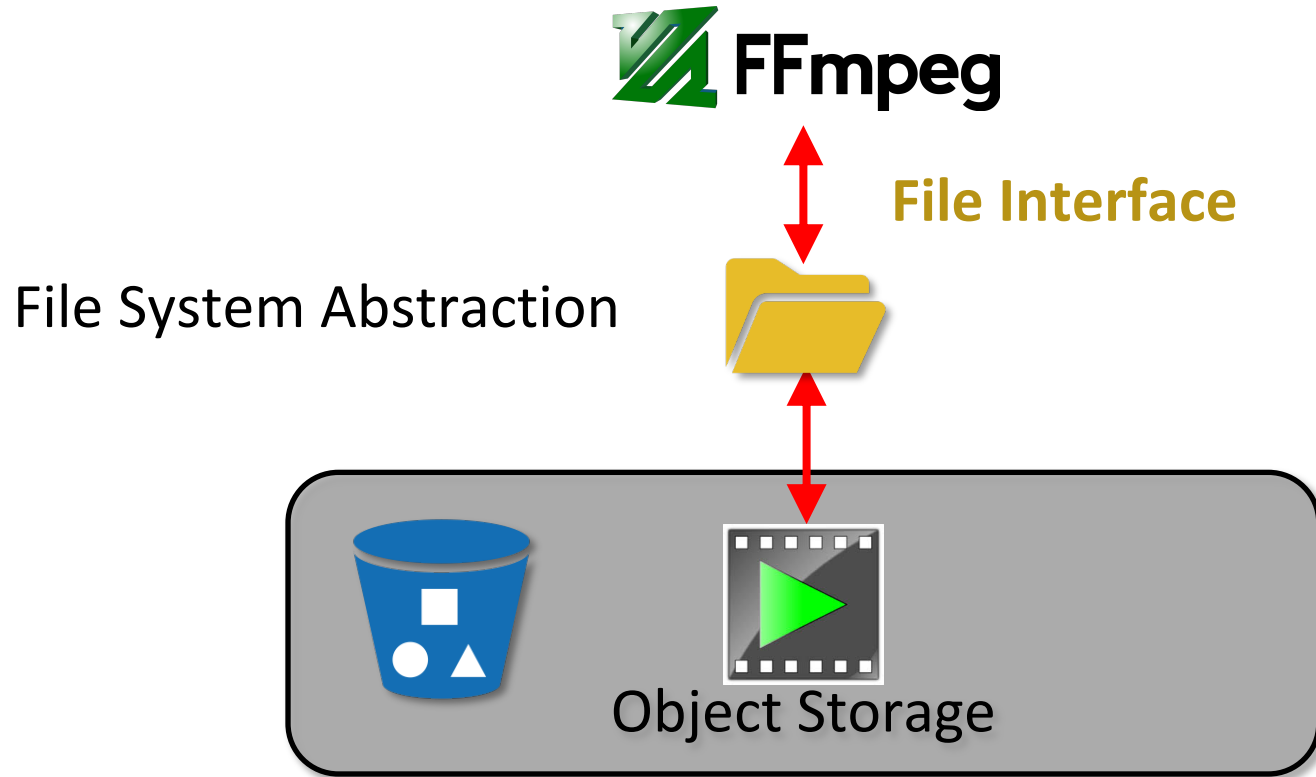
File Systems Paired With Object Storage



Object Storage File System



Object Storage File System



Existing Systems

Object Storage
File Systems

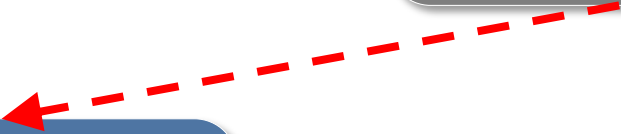


Existing Systems

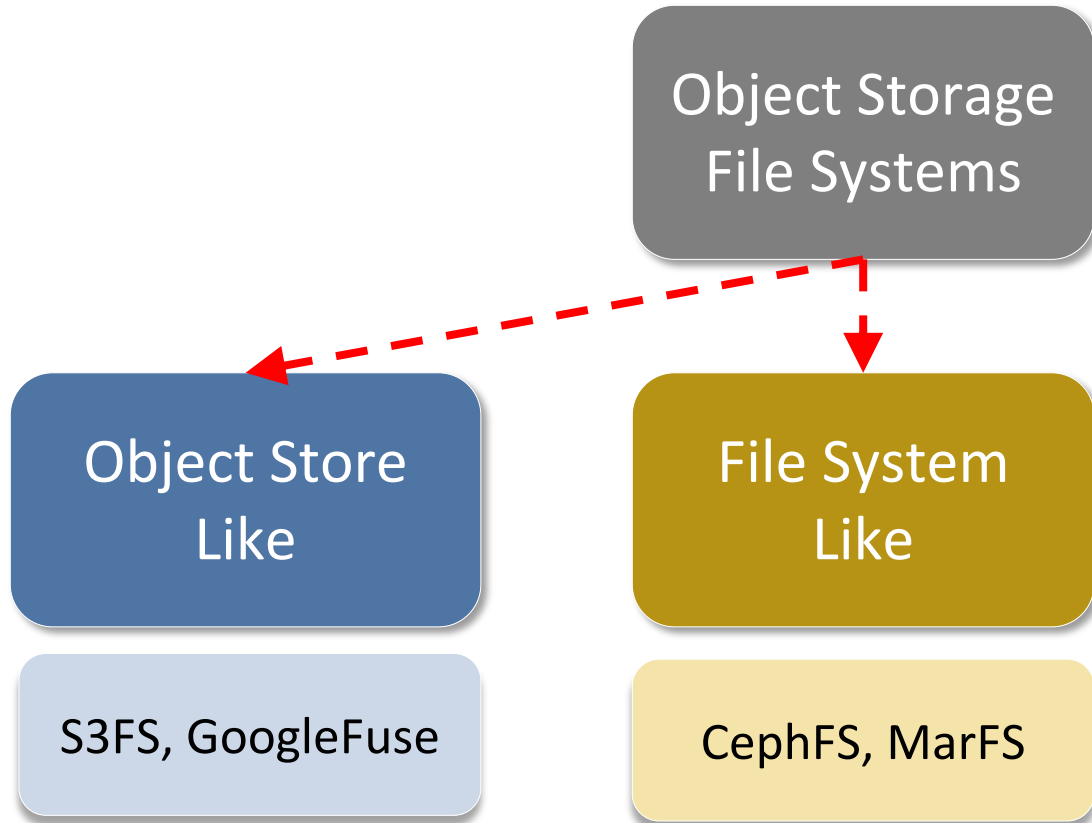
Object Storage
File Systems

Object Store
Like

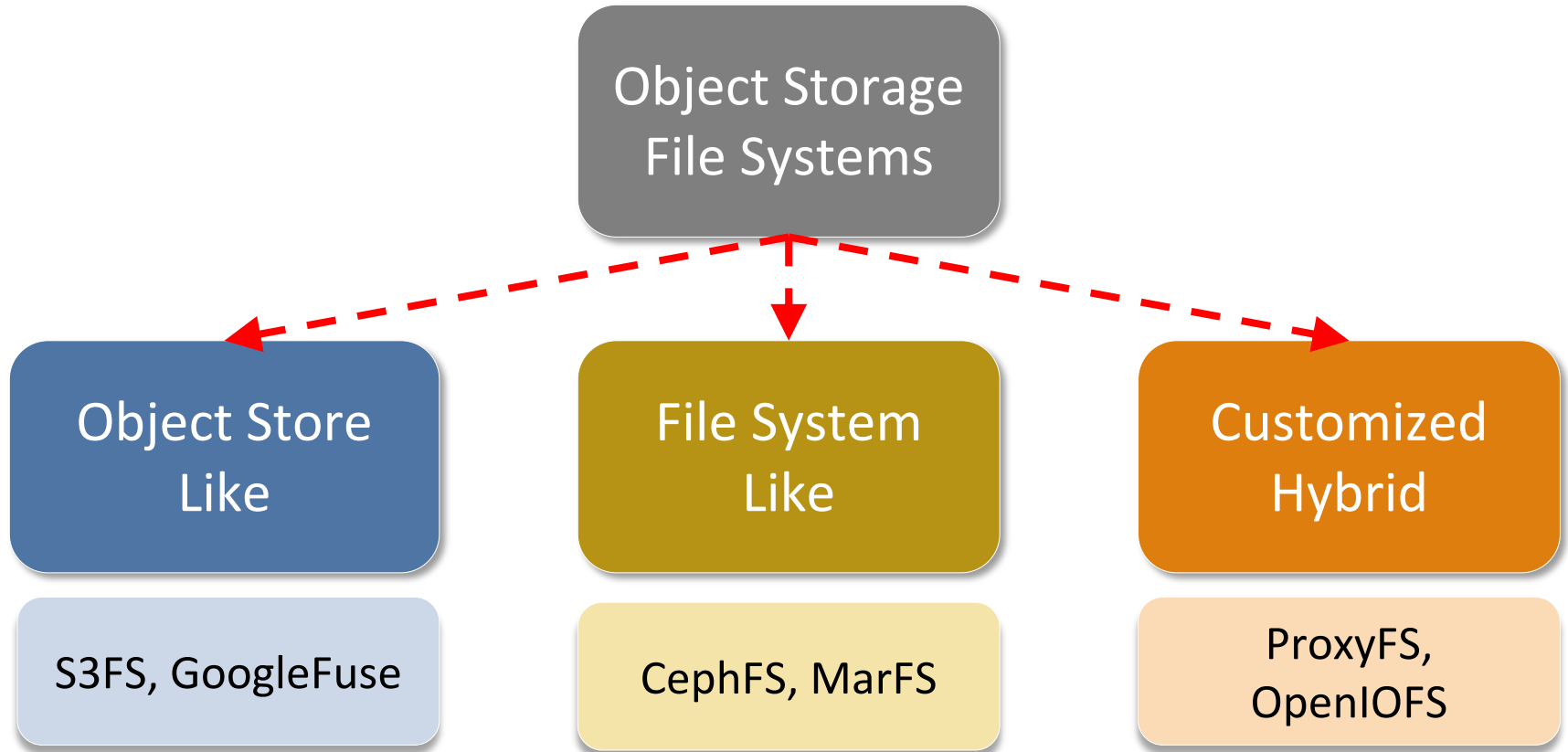
S3FS, GoogleFuse



Existing Systems

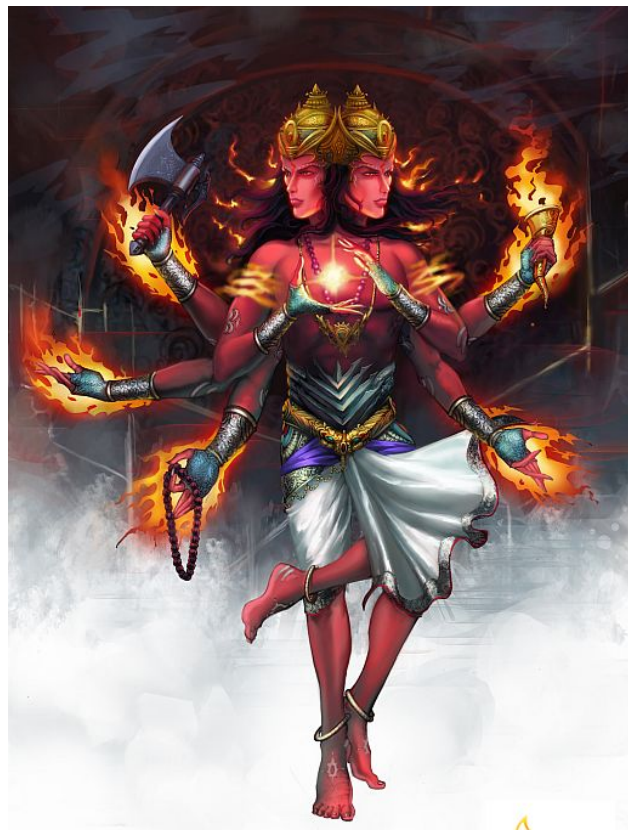


Existing Systems

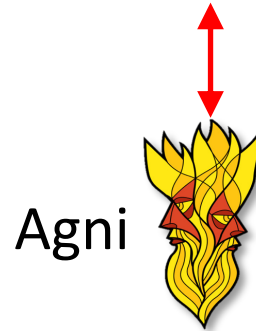


Outline

- ▶ Design considerations
- ▶ Existing systems
- ▶ Agni ←
- ▶ Vision



Architecture



Architecture



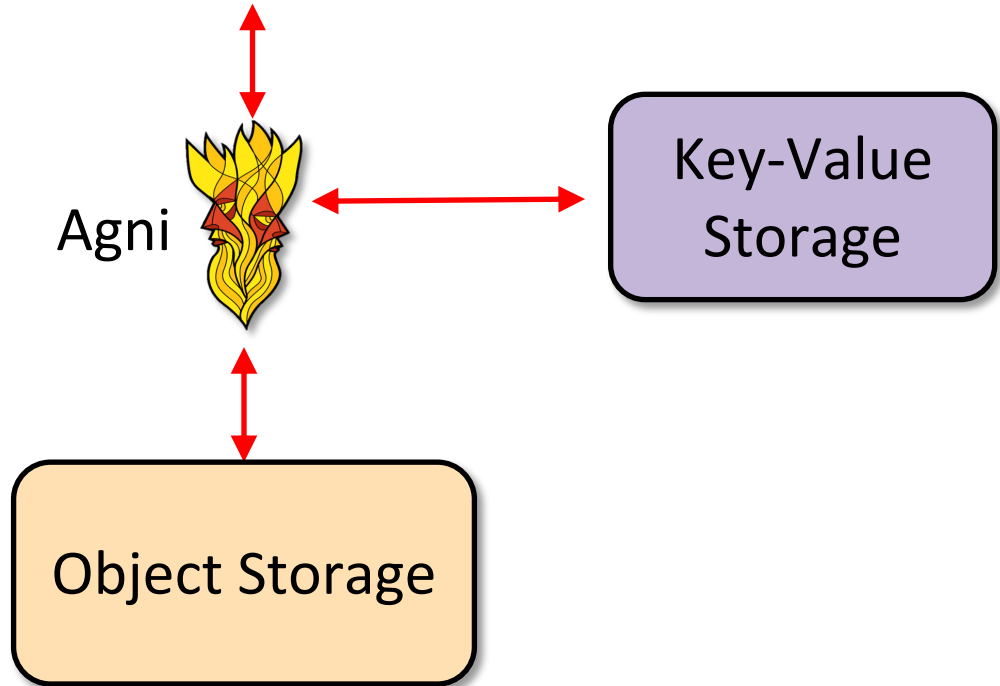
Agni



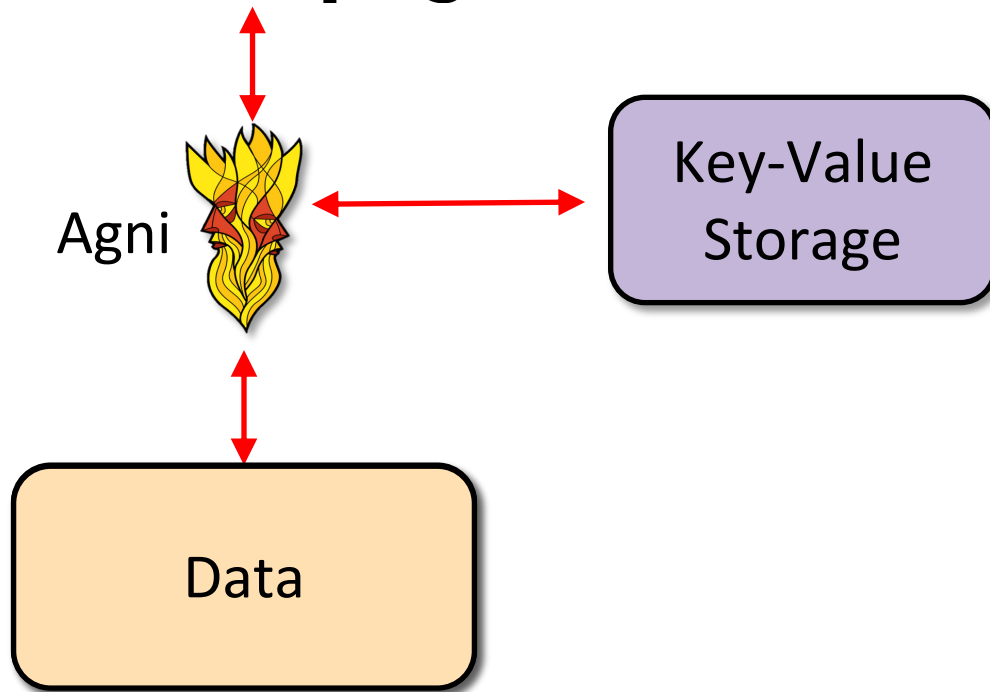
Object Storage



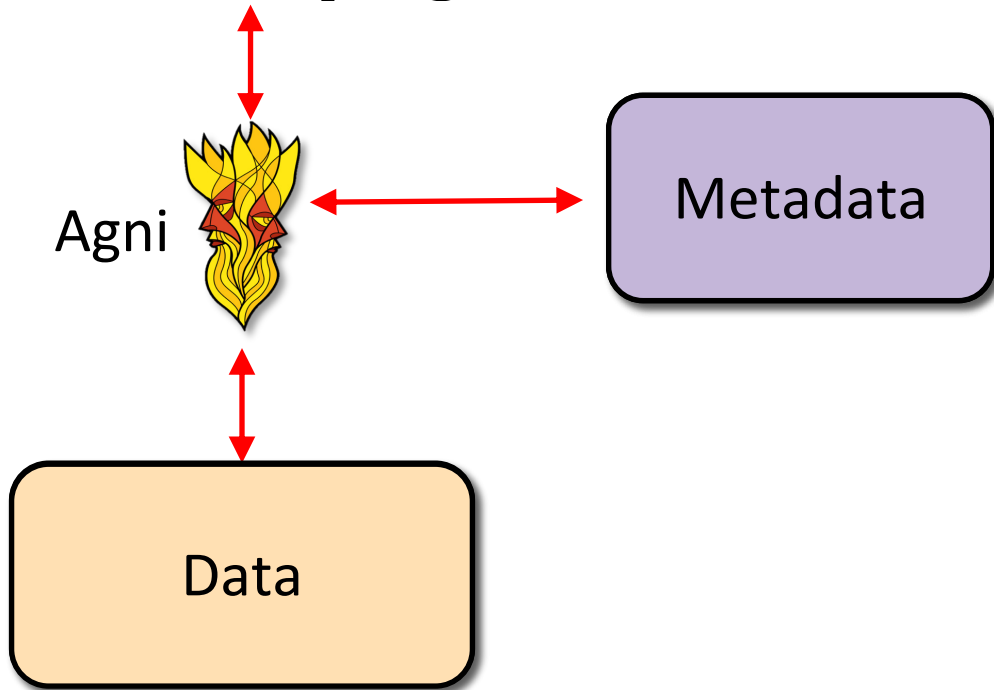
Architecture



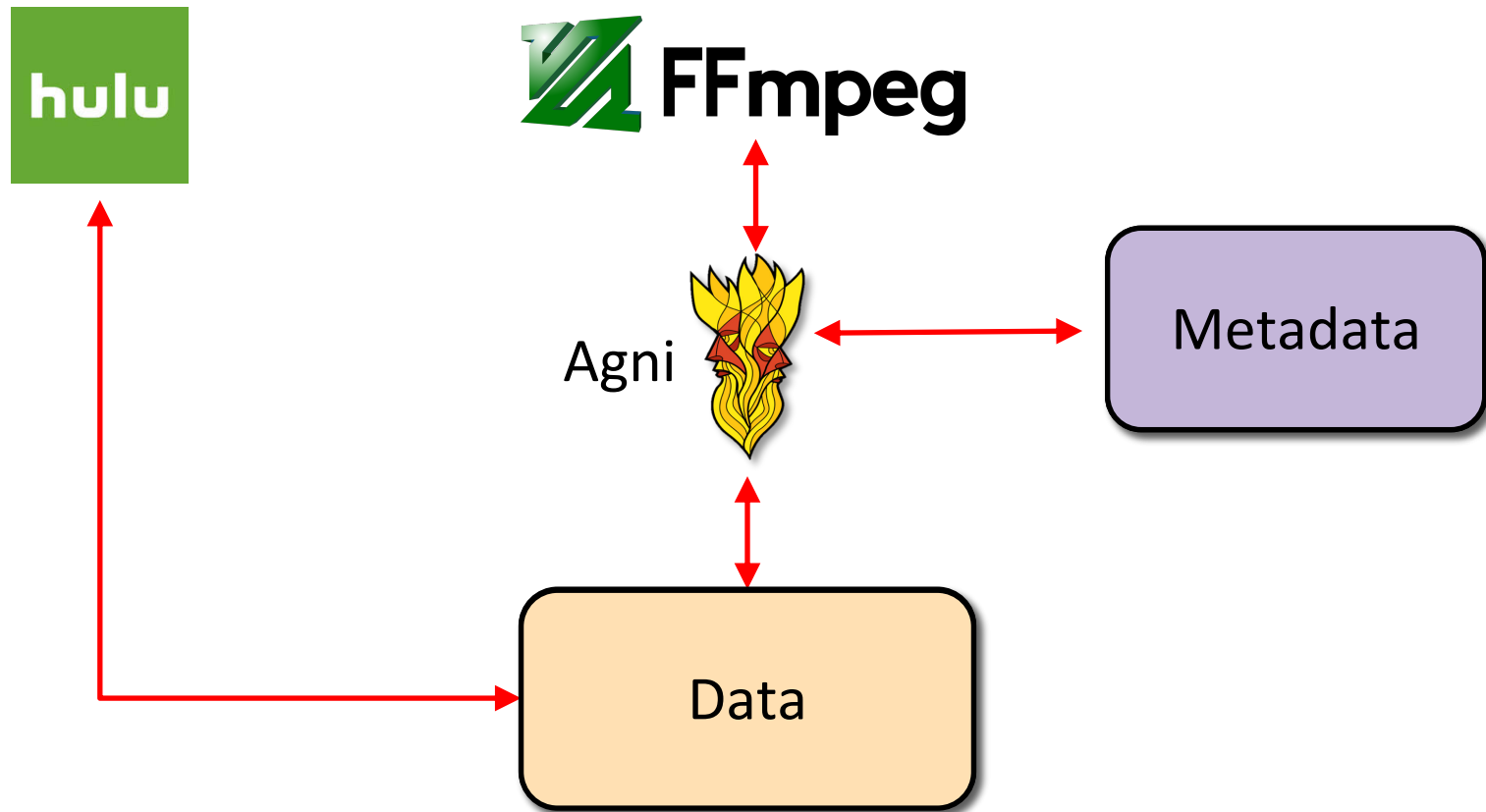
Architecture



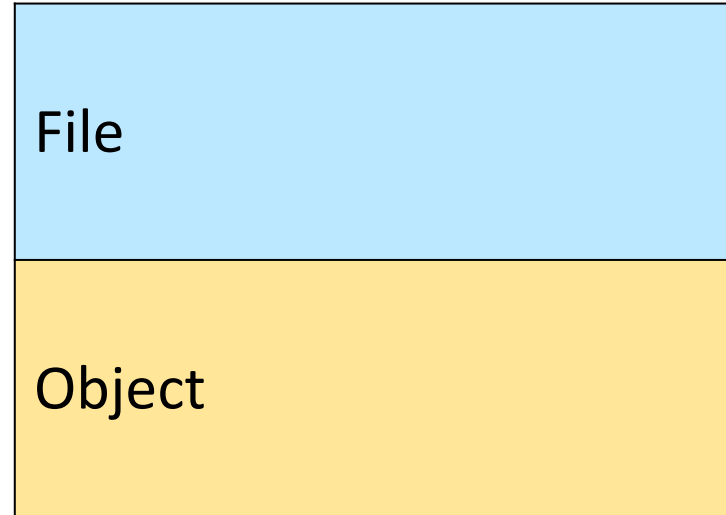
Architecture



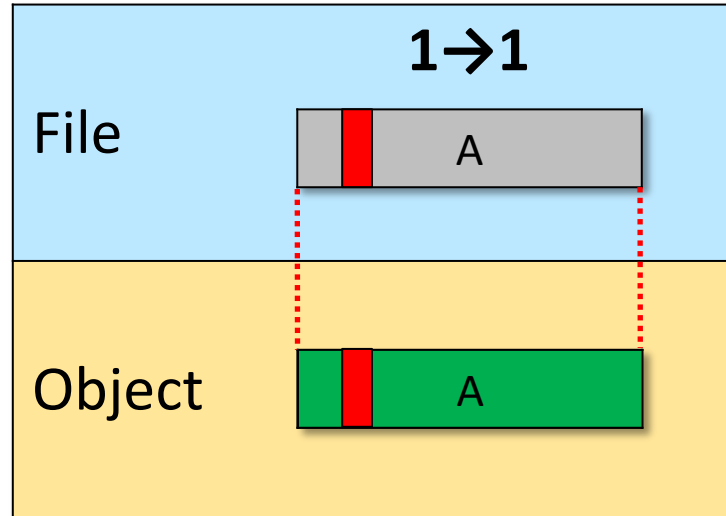
Architecture



Our Design Choices



Our Design Choices



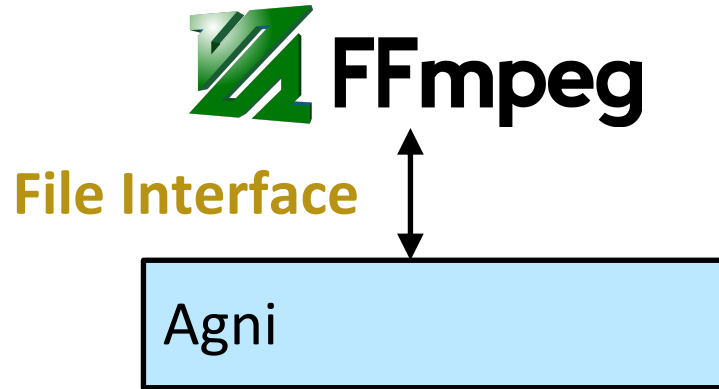
What about inefficient writes to immutable objects ?



Multi-Tier Data Structure



Multi-Tier Data Structure



Multi-Tier Data Structure



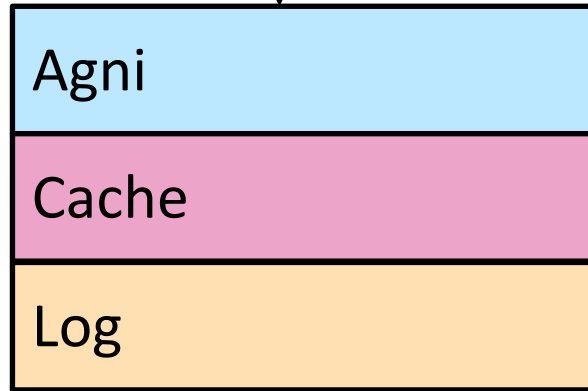
File Interface



Multi-Tier Data Structure



File Interface



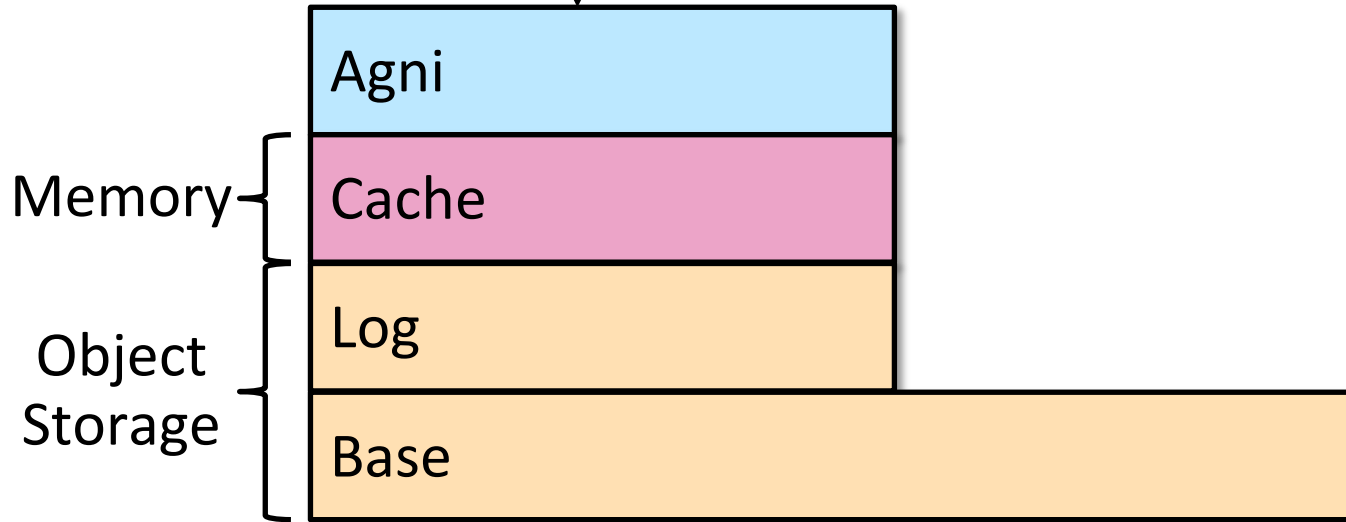
Memory



Multi-Tier Data Structure



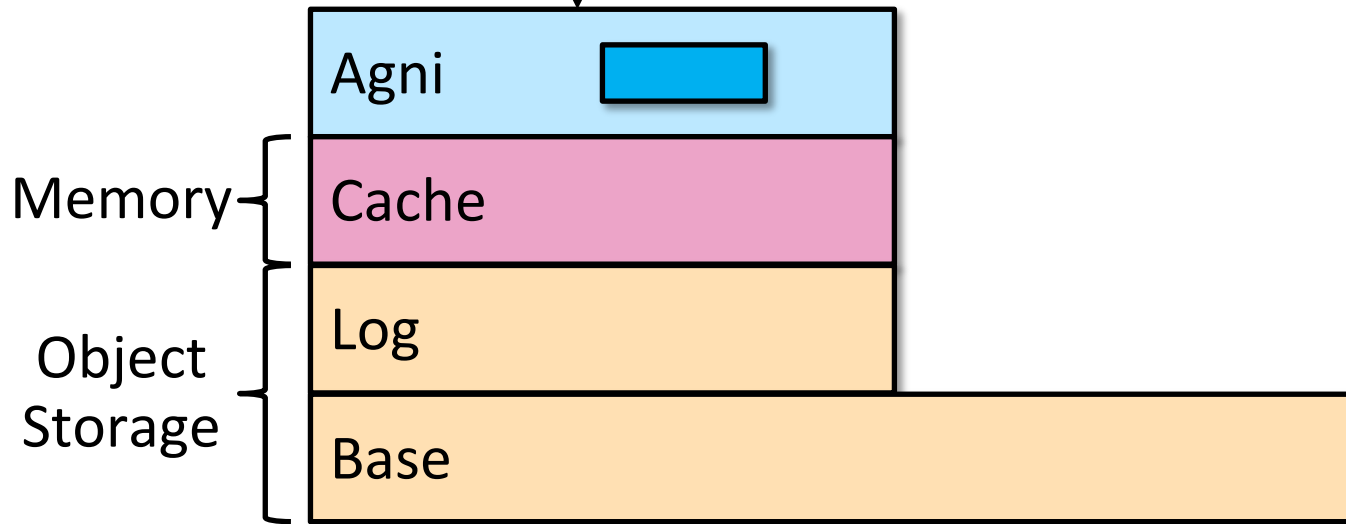
File Interface



Multi-Tier Data Structure



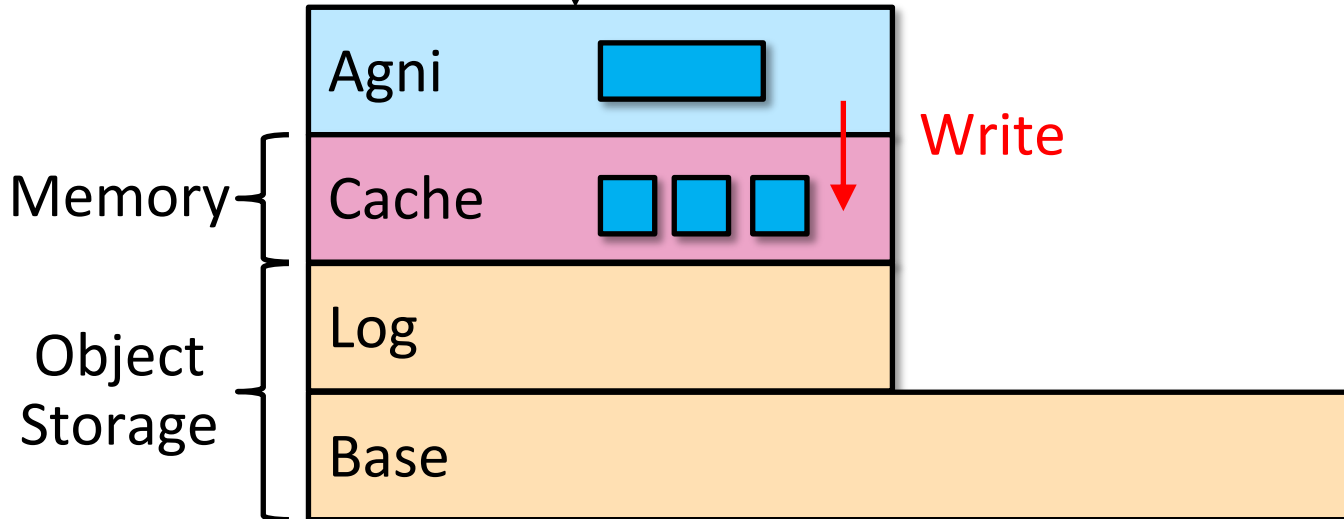
File Interface



Multi-Tier Data Structure



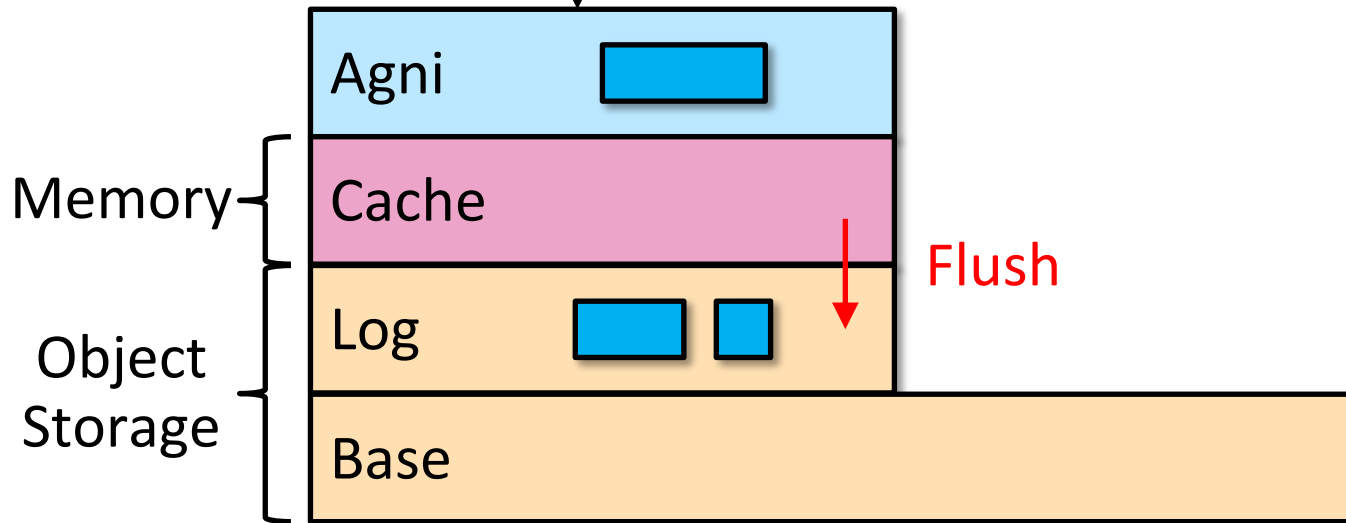
File Interface



Multi-Tier Data Structure



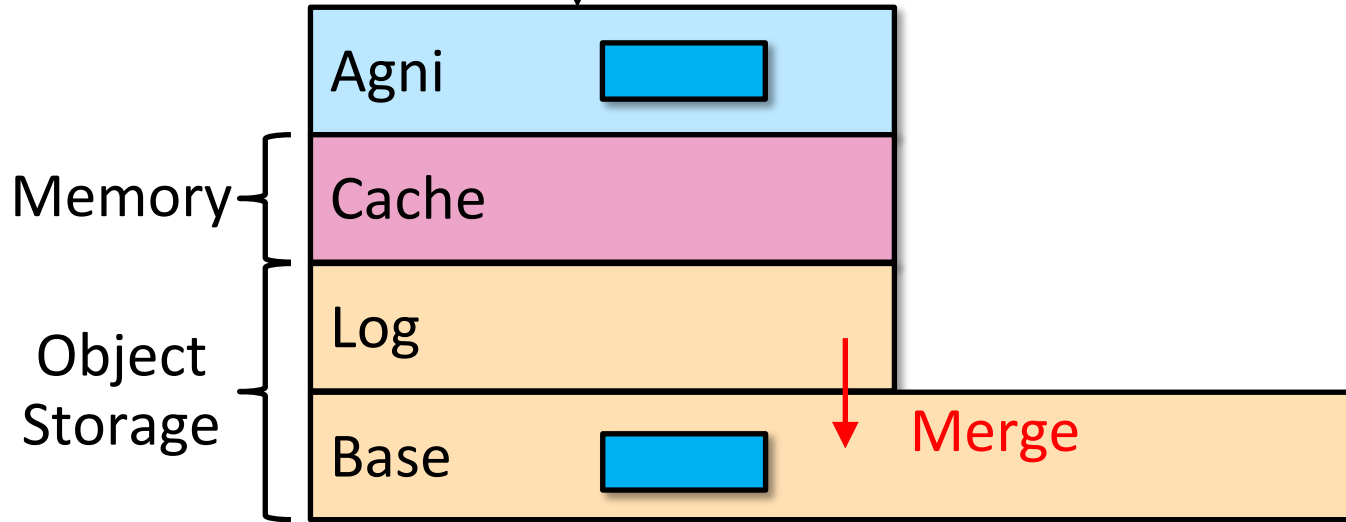
File Interface



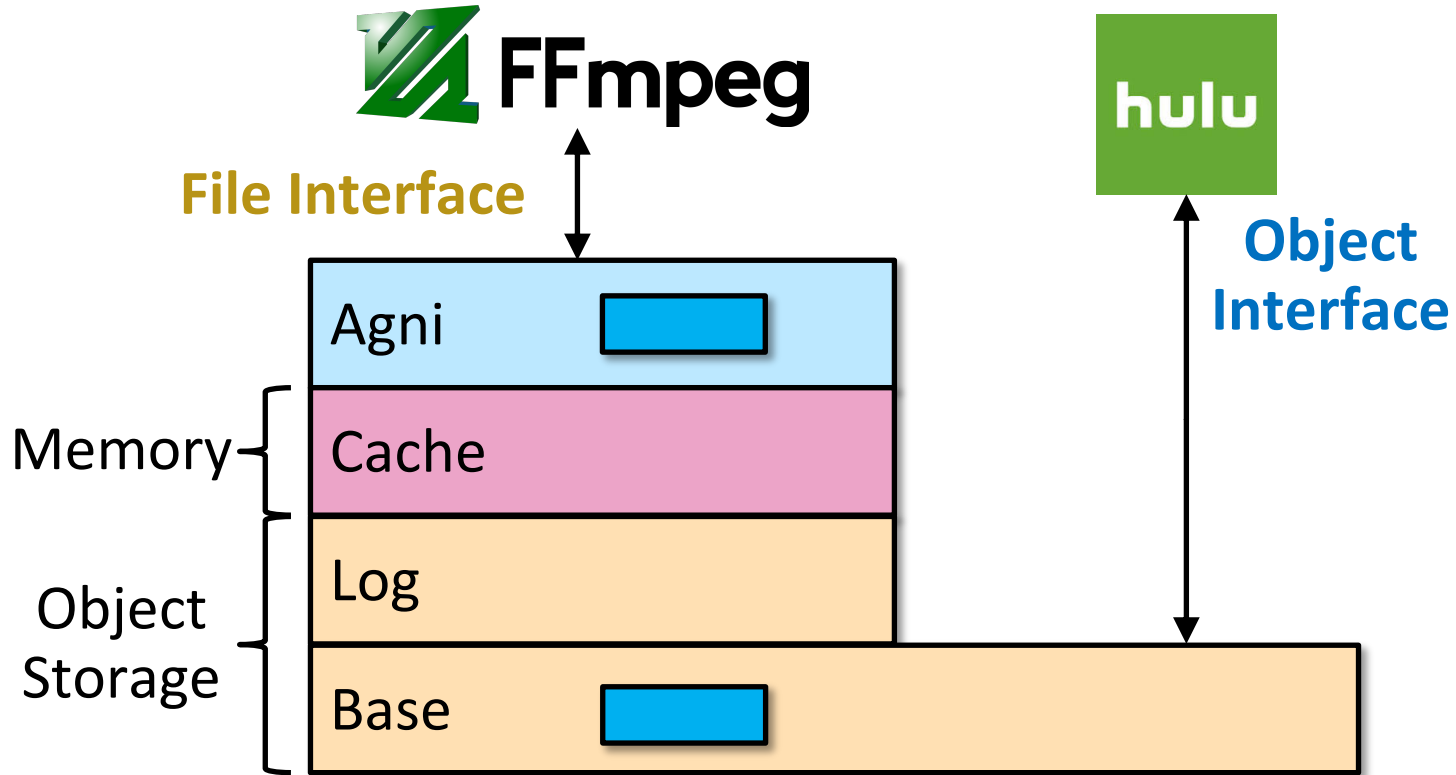
Multi-Tier Data Structure



File Interface



Multi-Tier Data Structure

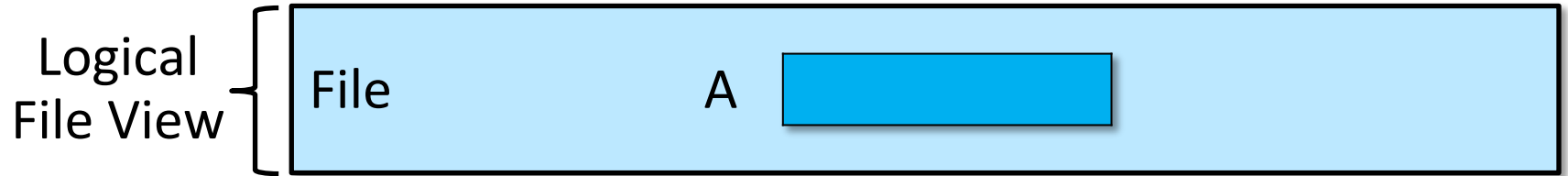


Data Layout

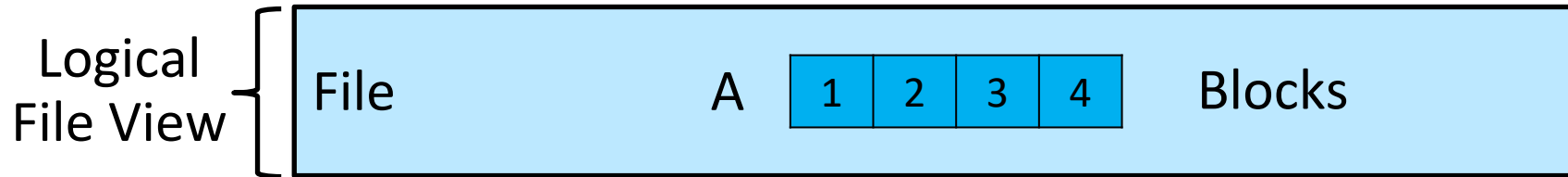
File



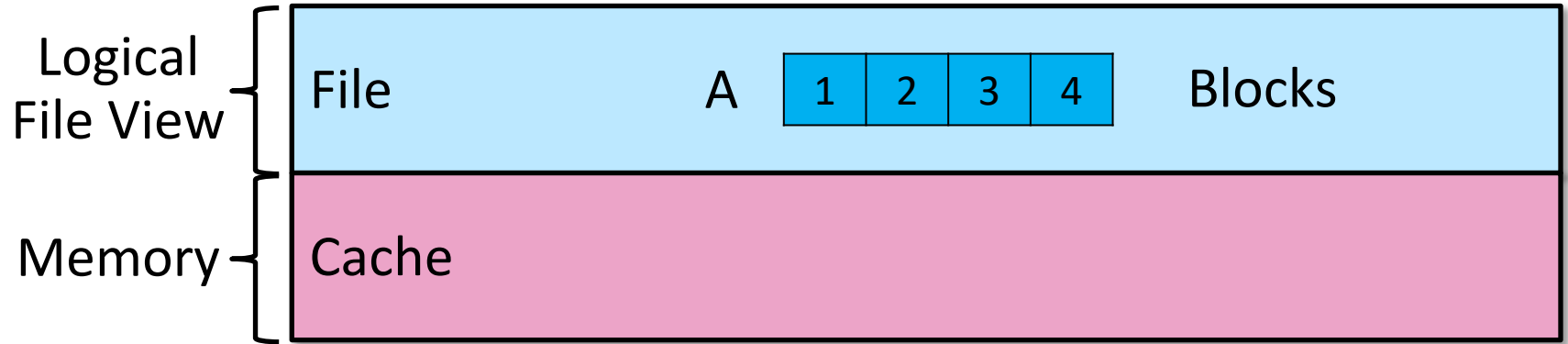
Data Layout



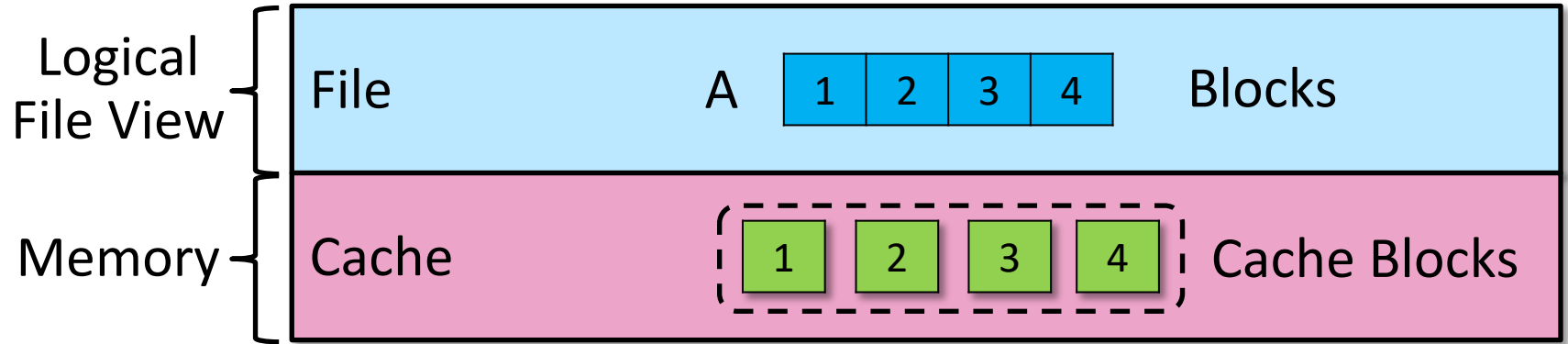
Data Layout



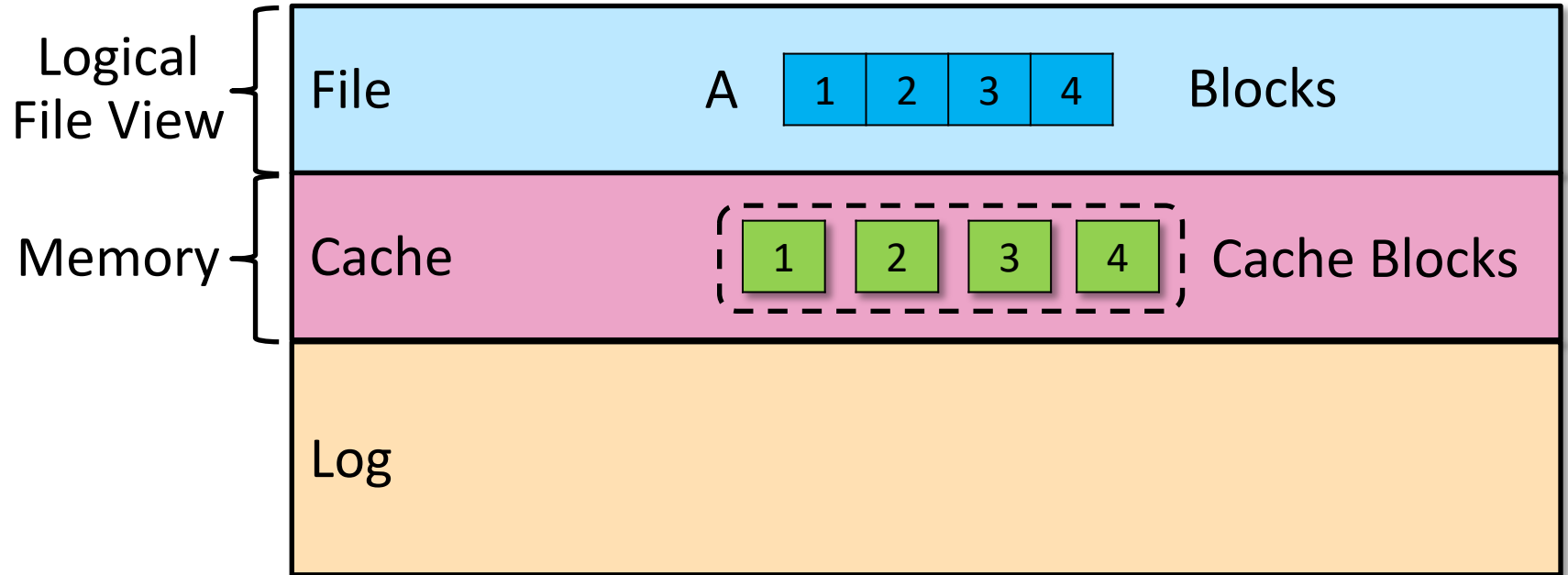
Data Layout



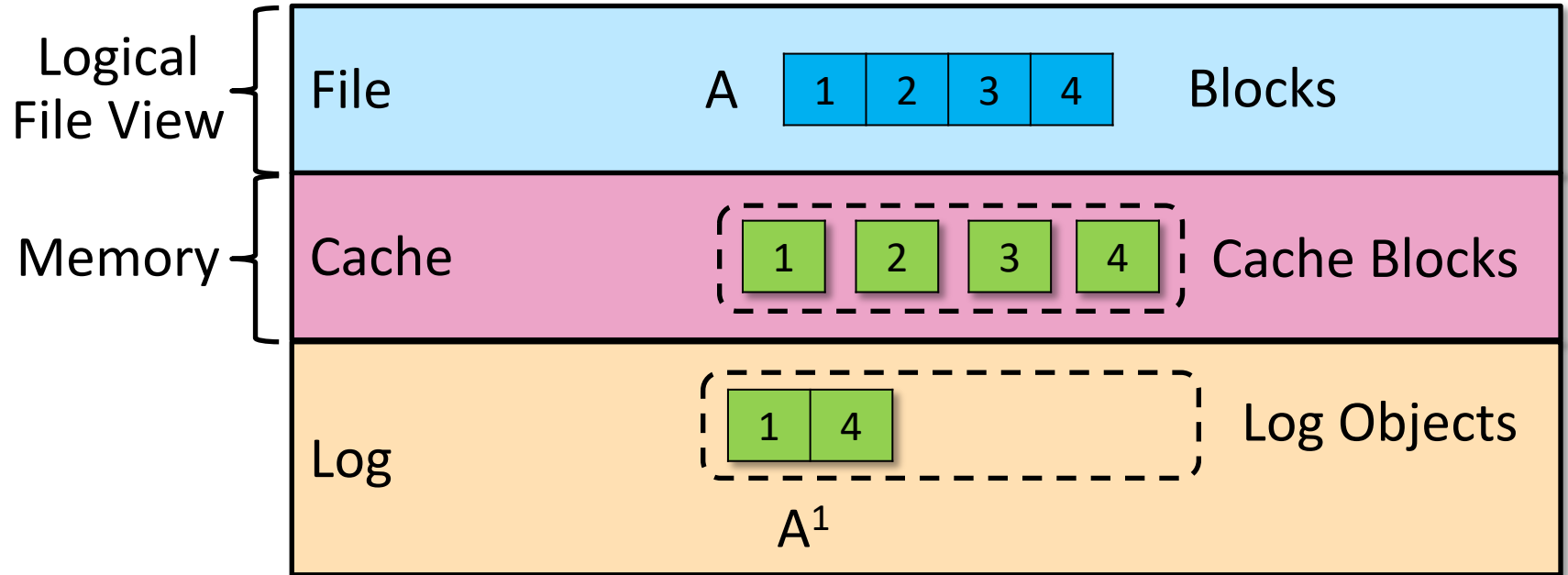
Data Layout



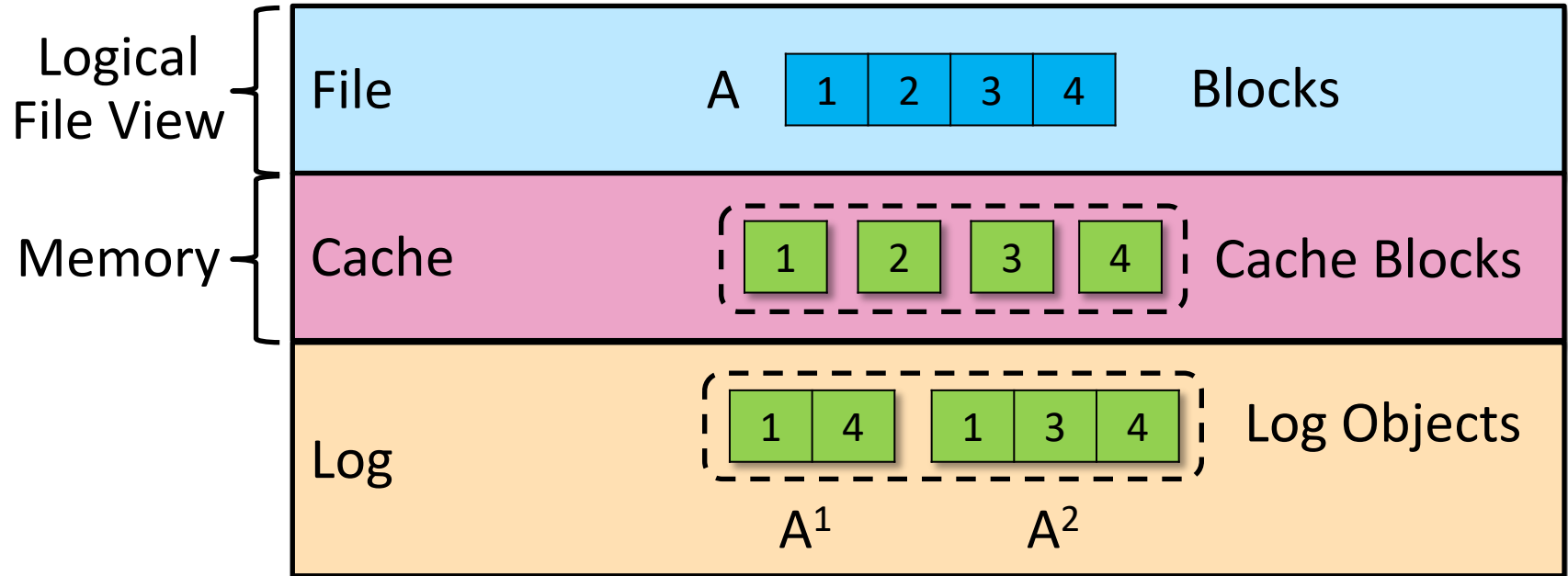
Data Layout



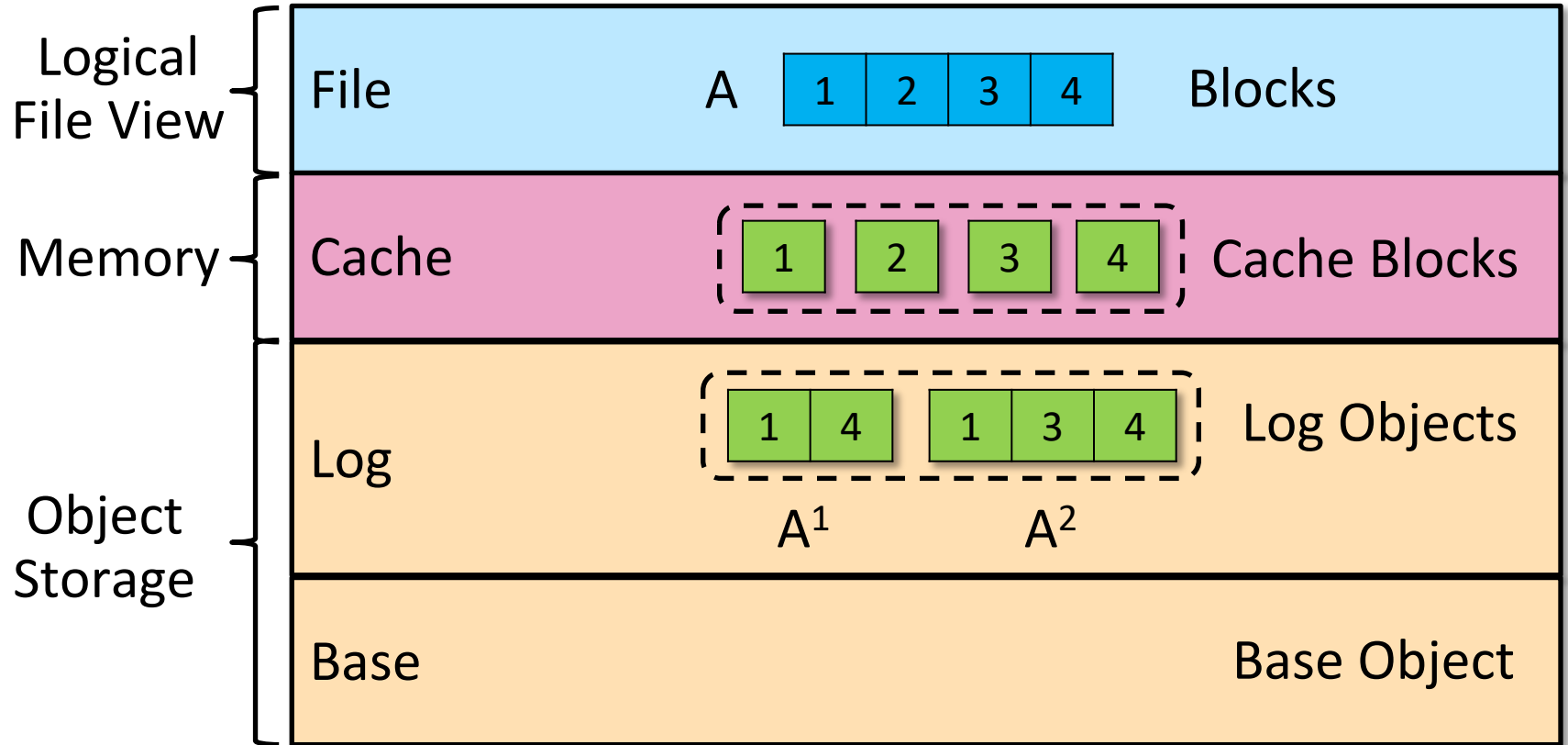
Data Layout



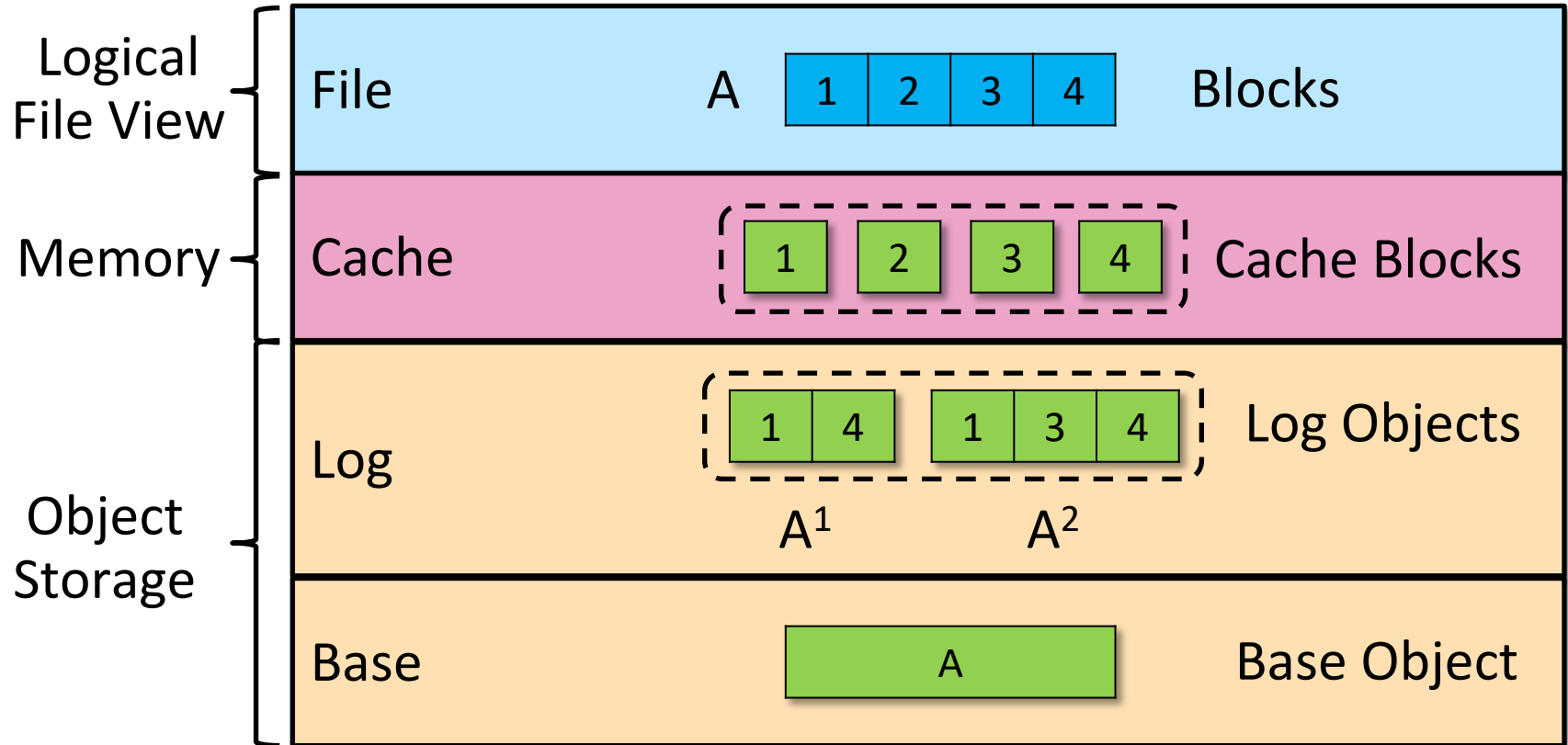
Data Layout



Data Layout



Data Layout



We Call This Approach **Eventual 1→1 Mapping**



File



Object



We Call This Approach **Eventual 1→1 Mapping**



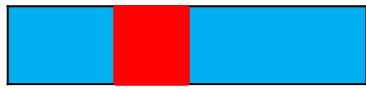
File



Object



We Call This Approach **Eventual 1→1 Mapping**



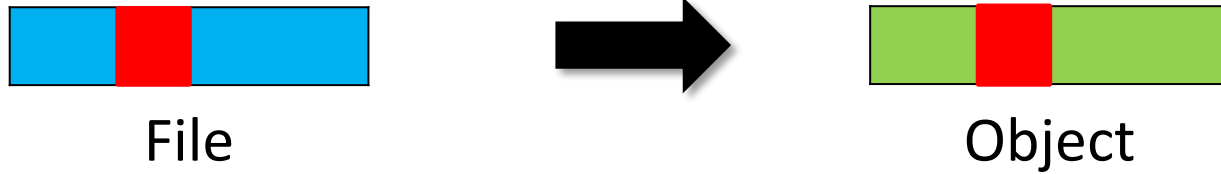
File



Object



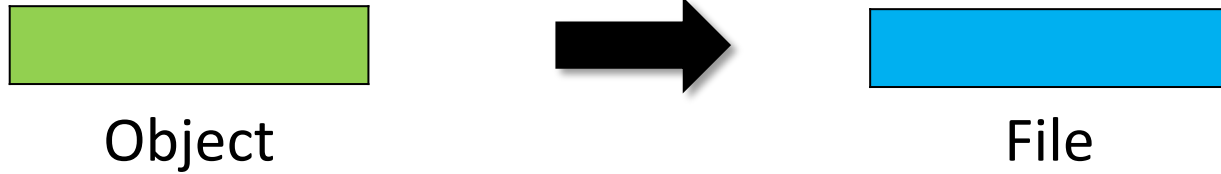
We Call This Approach **Eventual 1→1 Mapping**



File to Object Visibility Lag



Eventual 1→1 Mapping Works Both Ways



Eventual 1→1 Mapping Works Both Ways



Object



File



Eventual 1→1 Mapping Works Both Ways



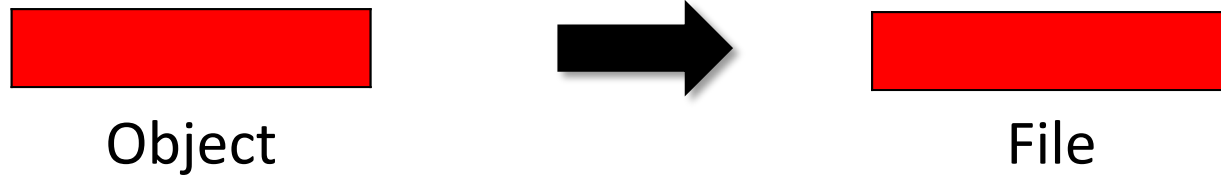
Object



File



Eventual 1→1 Mapping Works Both Ways



Object to File Visibility Lag



Design: Object to File



File Interface

Agni



Metadata

Data



Design: Object to File



Object
Interface

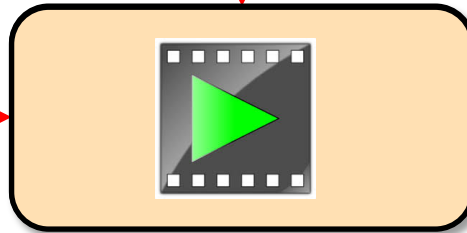


File Interface

Agni



Metadata



Design: Object to File



Object
Interface



File Interface

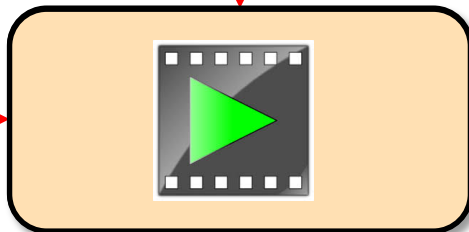
Agni



Visible

Metadata

Notification
Processor



Namespace Management



Namespace Management



Namespace Management

Master index

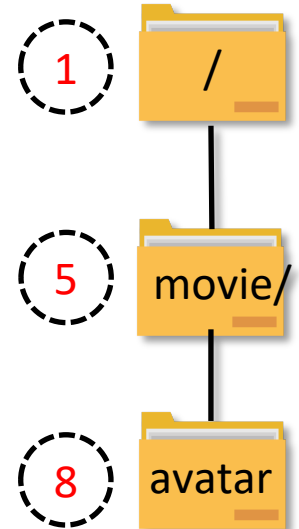
Key	Value



Namespace Management

Master index

Key	Value

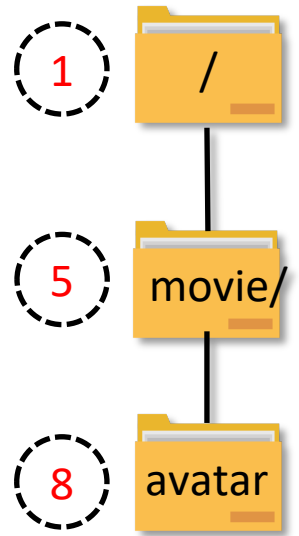


Namespace Management

Master index

Inode key

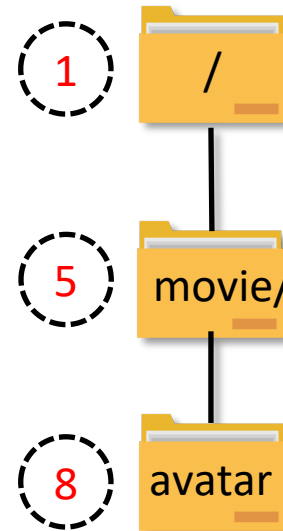
Key	Value
5	inode metadata



Namespace Management

Master index

	Key	Value
Inode key	5	inode metadata
Lookup key	1→movie	5

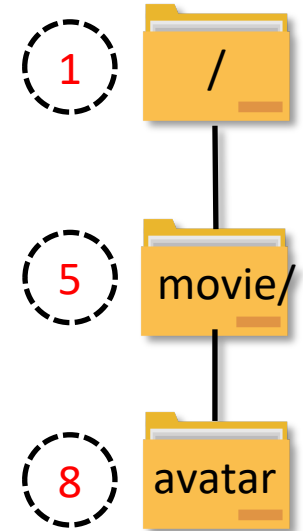


Namespace Management

Master index

	Key	Value
Inode key	5	inode metadata
Lookup key	1→movie	5
Children key	5→Children	[<8,avatar>]

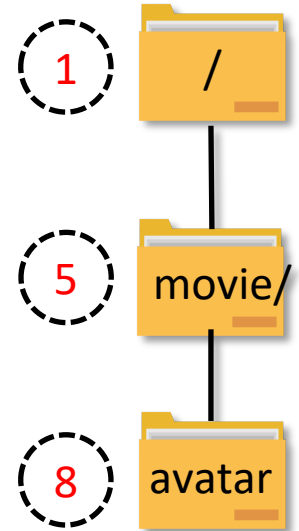
Directory metadata



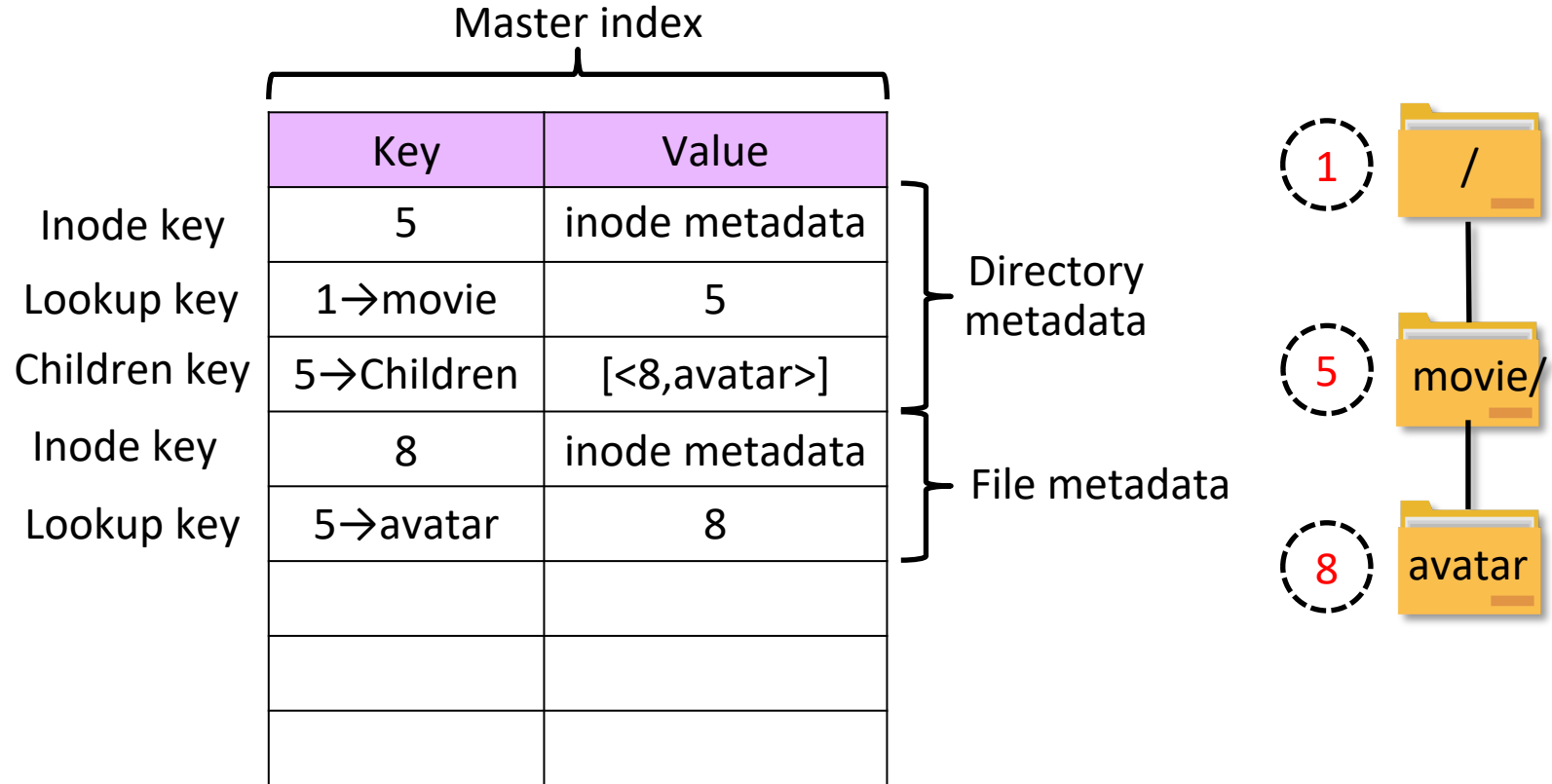
Namespace Management

Master index

	Key	Value	
Inode key	5	inode metadata	Directory metadata
Lookup key	1→movie	5	
Children key	5→Children	[<8,avatar>]	
Inode key	8	inode metadata	

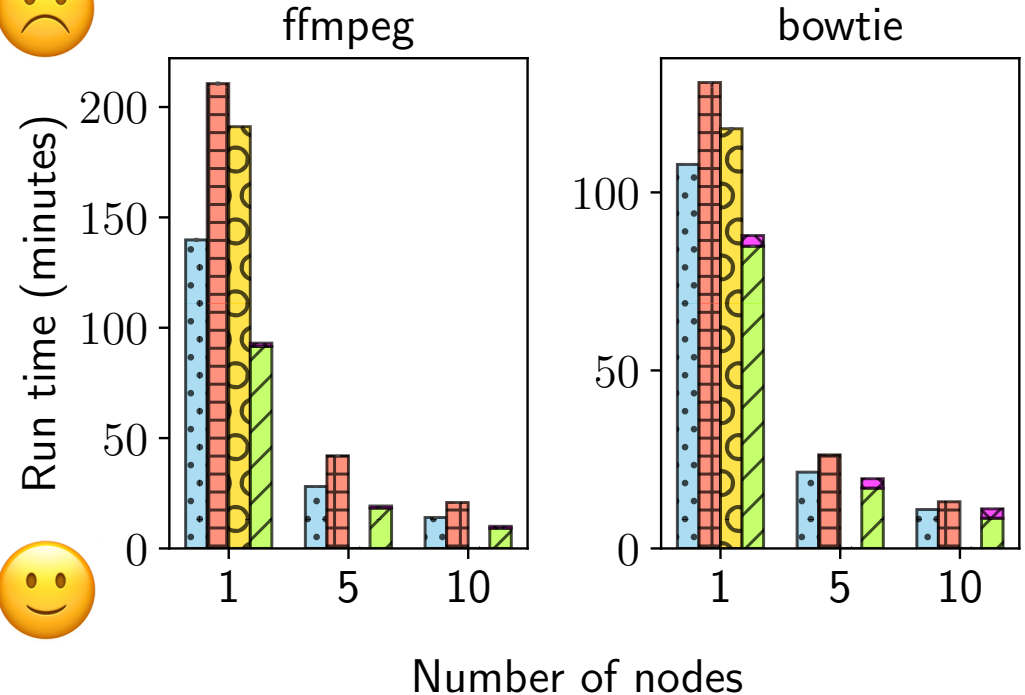
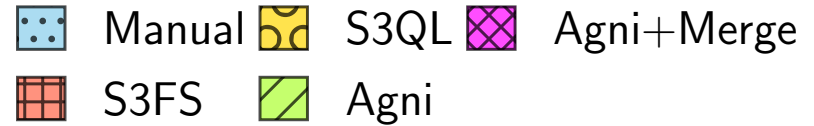


Namespace Management



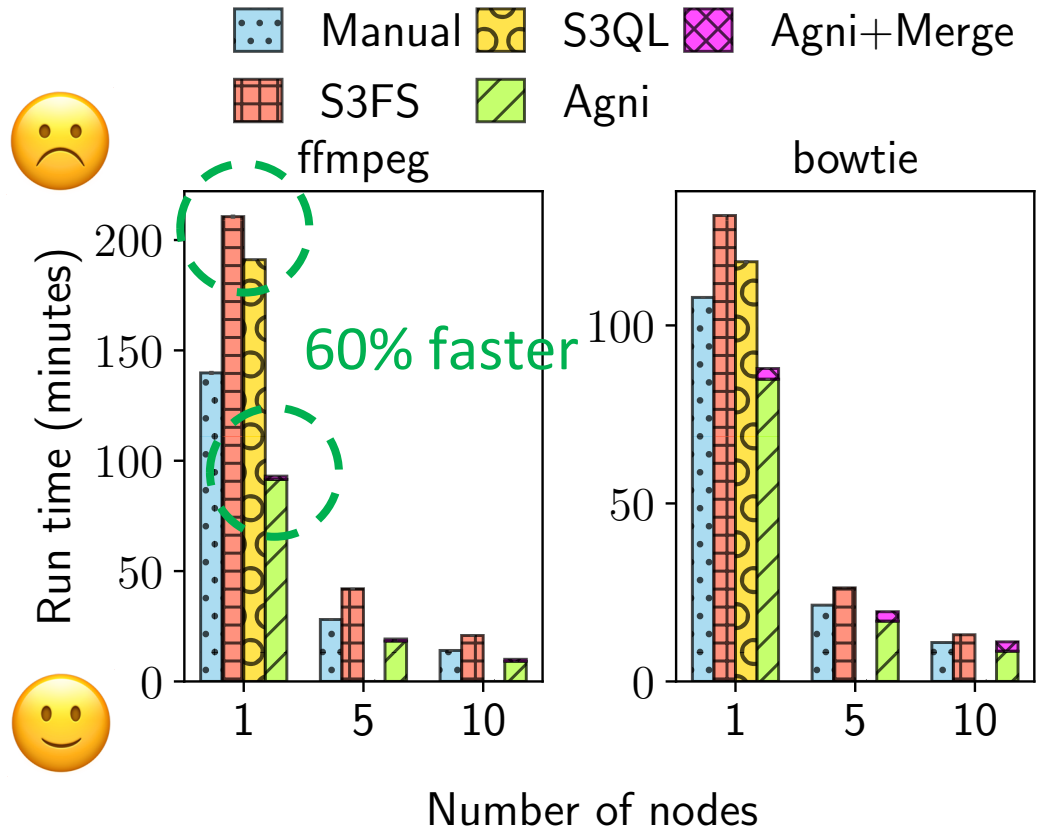
Evaluation: Applications

- ▶ **ffmpeg**: 320 GB of MPEG to MOV files 😞
- ▶ **bowtie**: 80 GB genome files
- ▶ Agni+Merge denotes when dual access is enabled 😊



Evaluation: Applications

- ▶ **ffmpeg**: 320 GB of MPEG to MOV files 😞
- ▶ **bowtie**: 80 GB genome files
- ▶ Agni+Merge denotes when dual access is enabled 😊



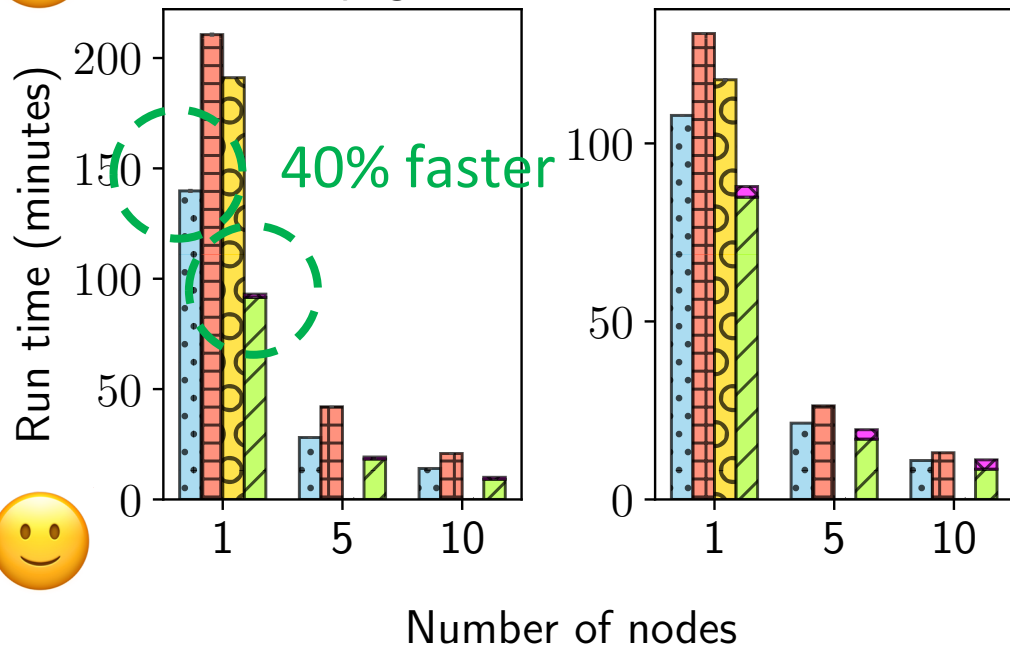
Evaluation: Applications

▶ **ffmpeg**: 320 GB of MPEG to MOV files



▶ **bowtie**: 80 GB genome files

▶ Agni+Merge denotes when dual access is enabled



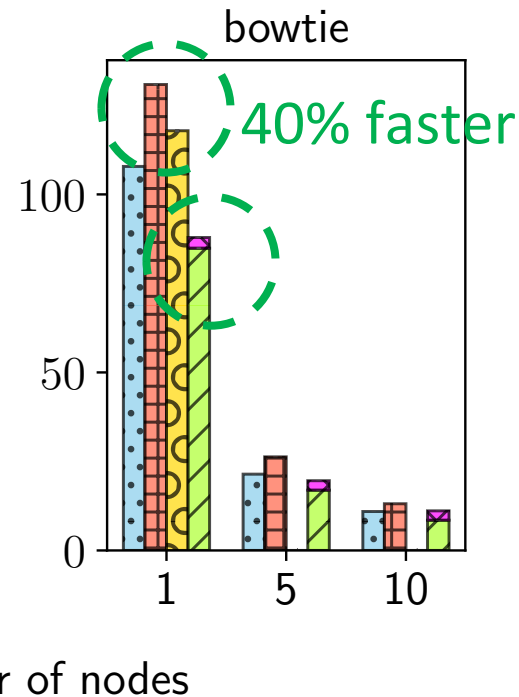
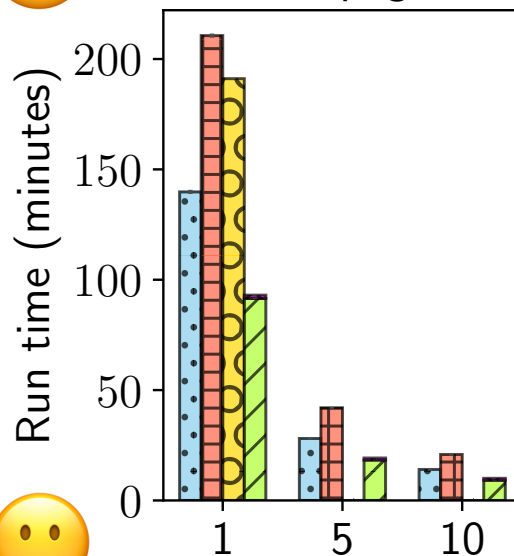
Evaluation: Applications

- ▶ **ffmpeg**: 320 GB of MPEG to MOV files



- ▶ **bowtie**: 80 GB genome files

- ▶ Agni+Merge denotes when dual access is enabled



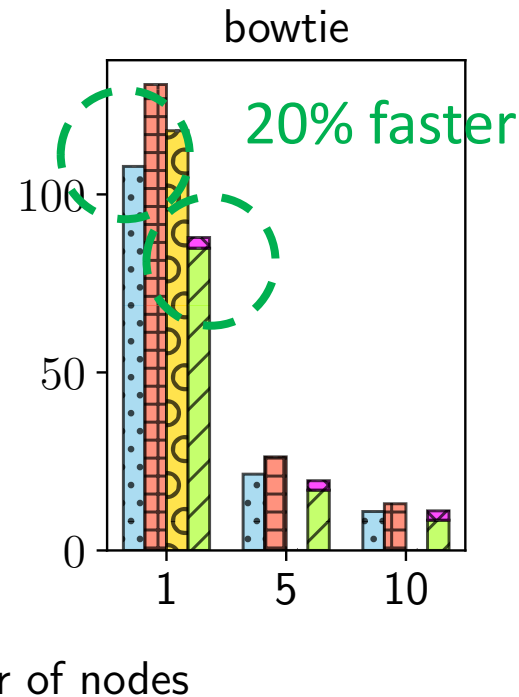
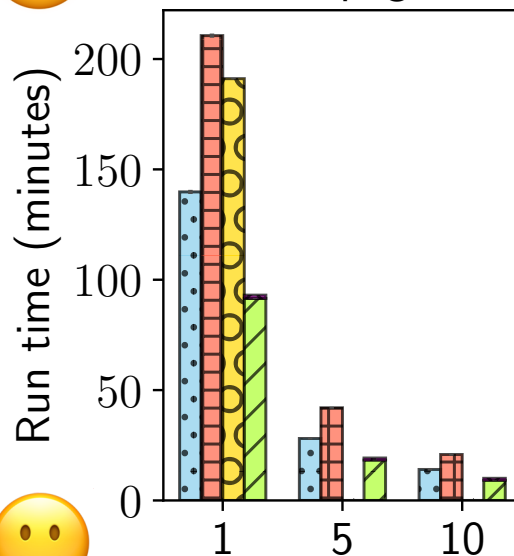
Evaluation: Applications

▶ **ffmpeg**: 320 GB of MPEG to MOV files



▶ **bowtie**: 80 GB genome files

▶ Agni+Merge denotes when dual access is enabled

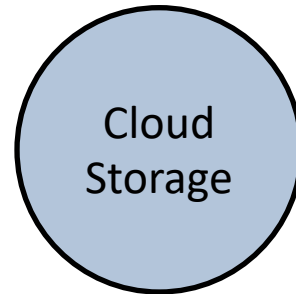


Outline

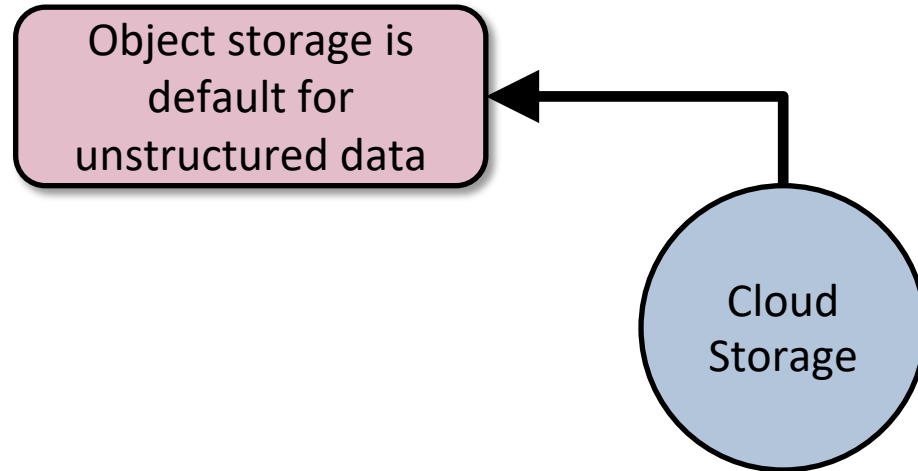
- ▶ Design considerations
- ▶ Existing systems
- ▶ Agni
- ▶ Vision



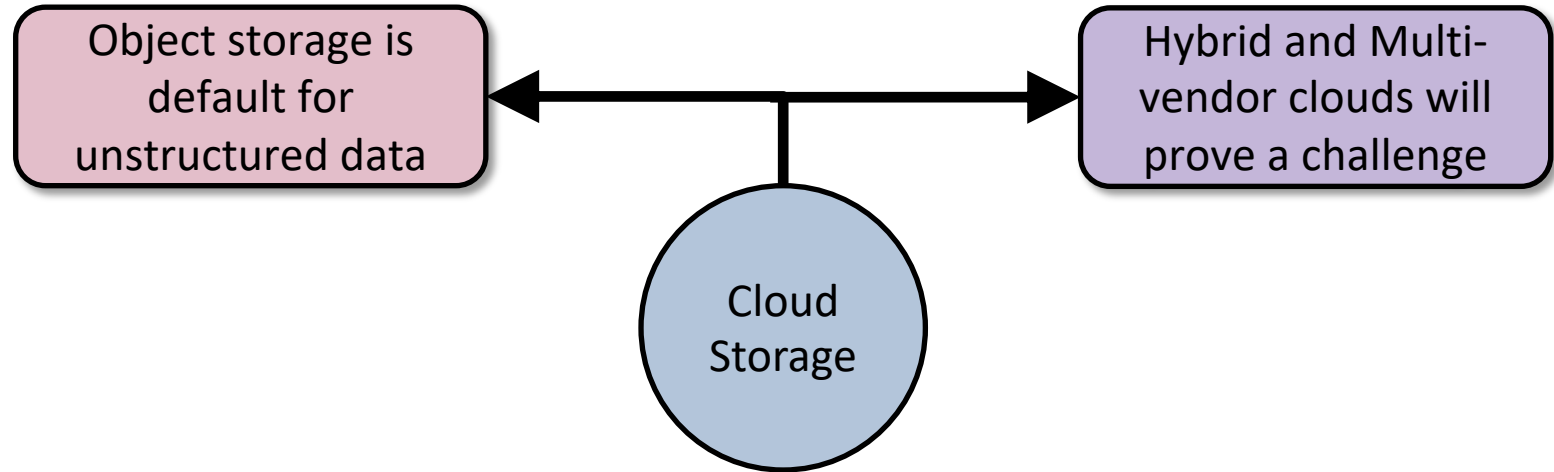
Vision for Cloud Storage



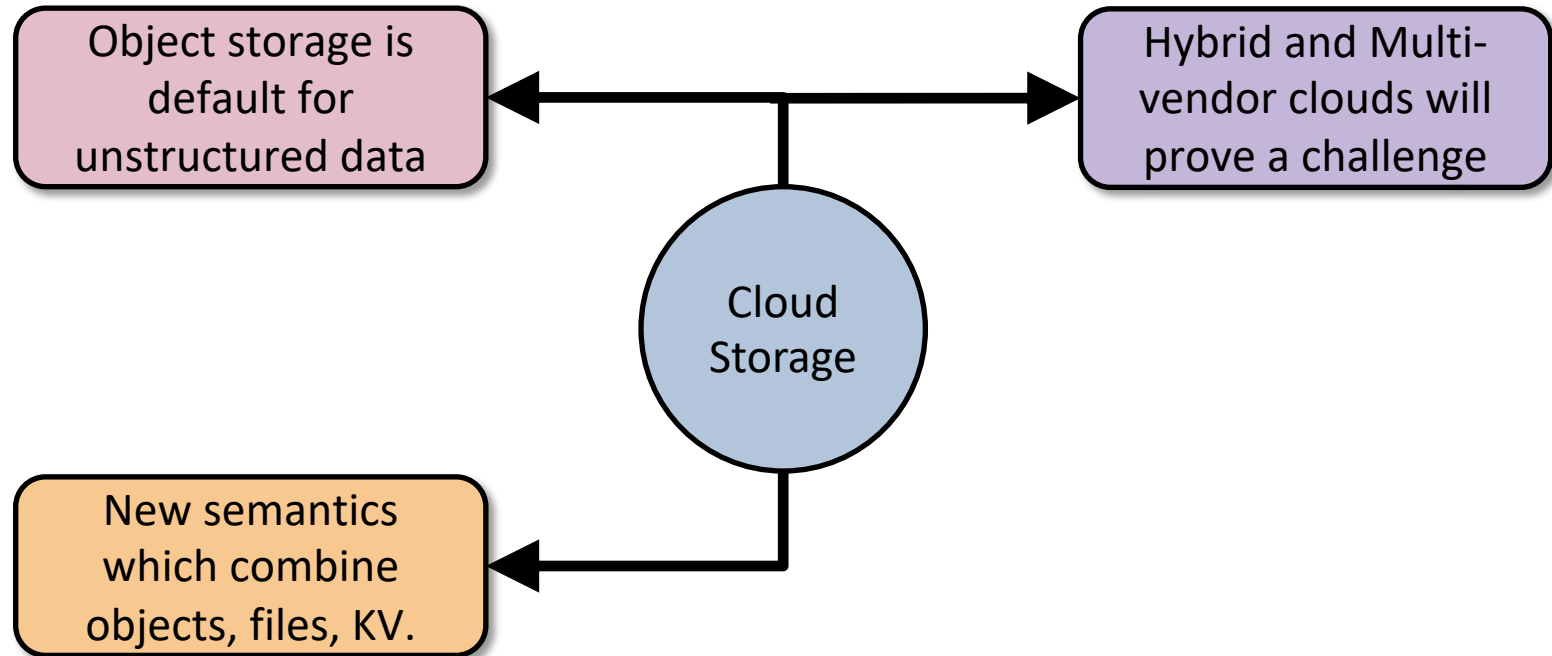
Vision for Cloud Storage



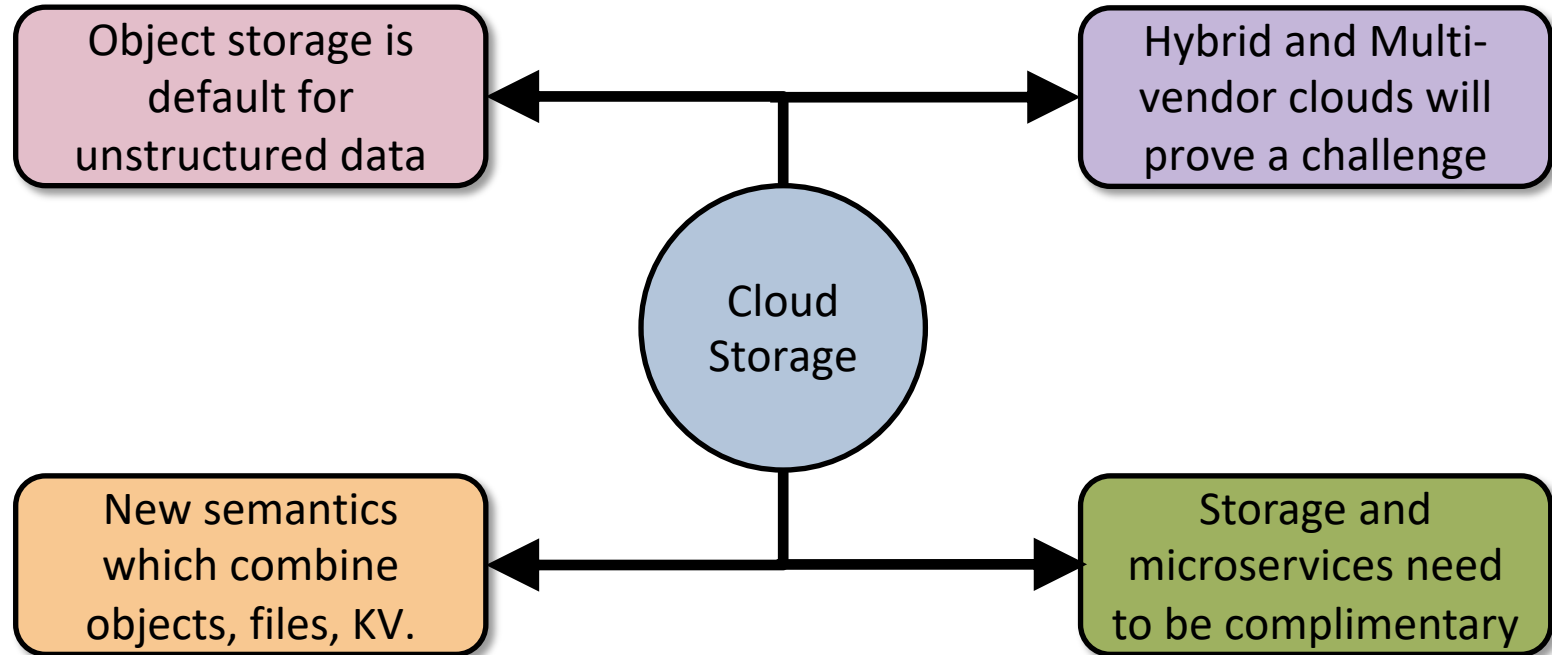
Vision for Cloud Storage



Vision for Cloud Storage



Vision for Cloud Storage





Thank you!

<https://github.com/objectfs/objectfs>

Contact: lillaney@jhu.edu

