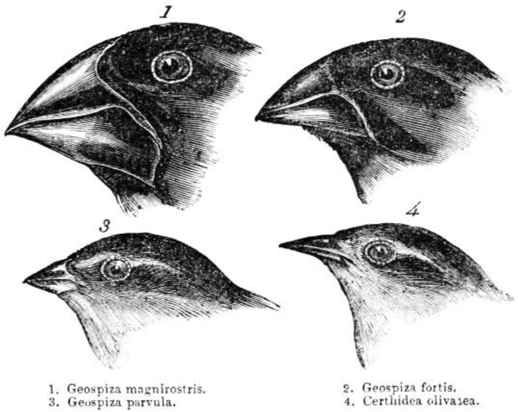
SOS 23 WORKSHOP

BYOFS:

THE OPPORTUNITIES AND DANGERS OF SPECIALIZATION IN THE AGE OF EXASCALE DATA STORAGE



PHILIP CARNS
Argonne National Laboratory

March 28, 2019
Asheville NC

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

WHAT DO YOU LIKE ABOUT YOUR PARALLEL FILE SYSTEM?

Or: what features spark joy?

Feature	User1	User2	Admin1
Random access read latency	***		
Large checkpoint throughput		***	
Access to Globus	***		
HDF5 support		**	
Quotas			***
Resilience	**		**

- Modern large-scale parallel file systems offer a wide range of sophisticated features! These are just some examples.

* User1, User2, and Admin1 are hypothetical characters.

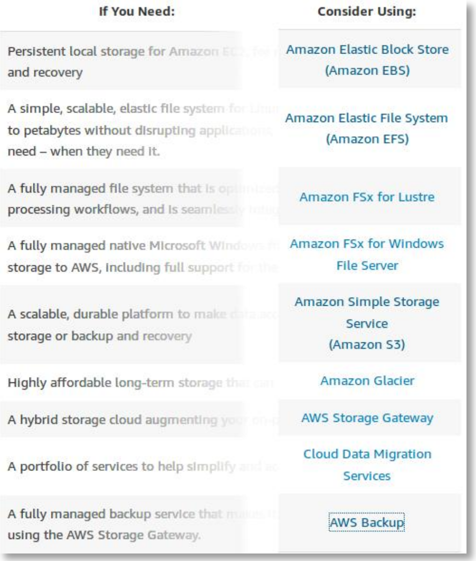
NOW THAT YOU KNOW WHAT YOU WANT…

- What storage choices are on the menu when you create an HPC account?
  1. A scratch file system, or maybe a few of them
  2. Project space
  3. A burst buffer (at some sites)
- You don’t actually choose those things, though, you just get them.
- And each is really just a deployment variation of a parallel file system.
  - different tradeoffs in capacity, performance, resilience, and availability
- Help yourself to your own accessories (high level libraries) if you want to customize.

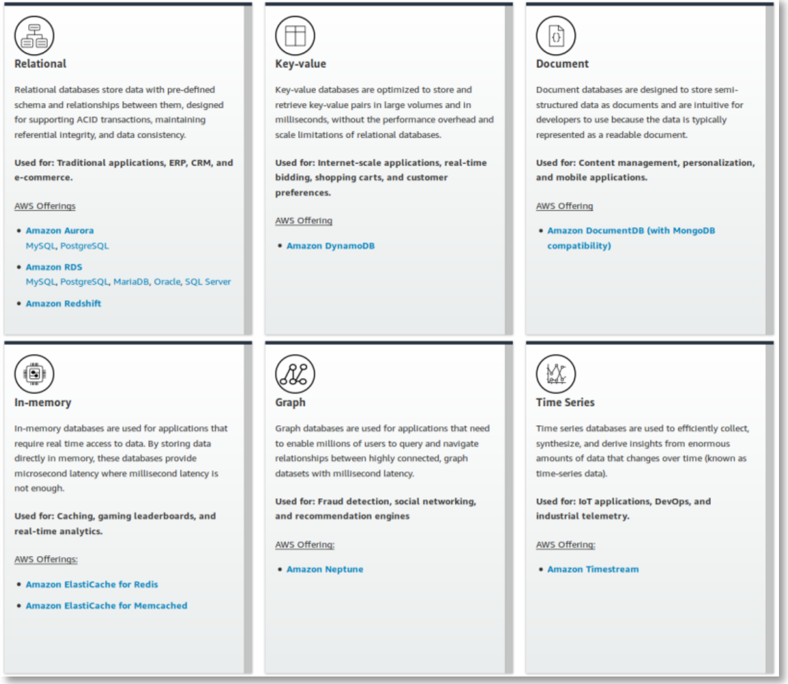
“*Any customer can have a car painted any color that he wants so long as it is black.*” -- H. Ford

IS THIS A UNIVERSAL PROBLEM?

Here are some of the storage options available when you sign up for an Amazon Web Services account:



If You Need:	Consider Using:
Persistent local storage for Amazon EC2, for and recovery	Amazon Elastic Block Store (Amazon EBS)
A simple, scalable, elastic file system for Linux to petabytes without disrupting applications, need – when they need it.	Amazon Elastic File System (Amazon EFS)
A fully managed file system that is optimized processing workflows, and is seamlessly integ	Amazon FSx for Lustre
A fully managed native Microsoft Windows fi storage to AWS, including full support for the	Amazon FSx for Windows File Server
A scalable, durable platform to make data acc storage or backup and recovery	Amazon Simple Storage Service (Amazon S3)
Highly affordable long-term storage that can	Amazon Glacier
A hybrid storage cloud augmenting your on-p	AWS Storage Gateway
A portfolio of services to help simplify and ac	Cloud Data Migration Services
A fully managed backup service that makes it using the AWS Storage Gateway.	AWS Backup



Relational

Relational databases store data with pre-defined schema and relationships between them, designed for supporting ACID transactions, maintaining referential integrity, and data consistency.

Used for: Traditional applications, ERP, CRM, and e-commerce.

AWS Offerings

- **Amazon Aurora**
  MySQL, PostgreSQL
- **Amazon RDS**
  MySQL, PostgreSQL, MariaDB, Oracle, SQL Server
- **Amazon Redshift**

Key-value

Key-value databases are optimized to store and retrieve key-value pairs in large volumes and in milliseconds, without the performance overhead and scale limitations of relational databases.

Used for: Internet-scale applications, real-time bidding, shopping carts, and customer preferences.

AWS Offering

- **Amazon DynamoDB**

Document

Document databases are designed to store semi-structured data as documents and are intuitive for developers to use because the data is typically represented as a readable document.

Used for: Content management, personalization, and mobile applications.

AWS Offering

- **Amazon DocumentDB (with MongoDB compatibility)**

In-memory

In-memory databases are used for applications that require real time access to data. By storing data directly in memory, these databases provide microsecond latency where millisecond latency is not enough.

Used for: Caching, gaming leaderboards, and real-time analytics.

AWS Offerings:

- **Amazon ElastiCache for Redis**
- **Amazon ElastiCache for Memcached**

Graph

Graph databases are used for applications that need to enable millions of users to query and navigate relationships between highly connected, graph datasets with millisecond latency.

Used for: Fraud detection, social networking, and recommendation engines

AWS Offering:

- **Amazon Neptune**

Time Series

Time series databases are used to efficiently collect, synthesize, and derive insights from enormous amounts of data that changes over time (known as time-series data).

Used for: IoT applications, DevOps, and industrial telemetry.

AWS Offering:

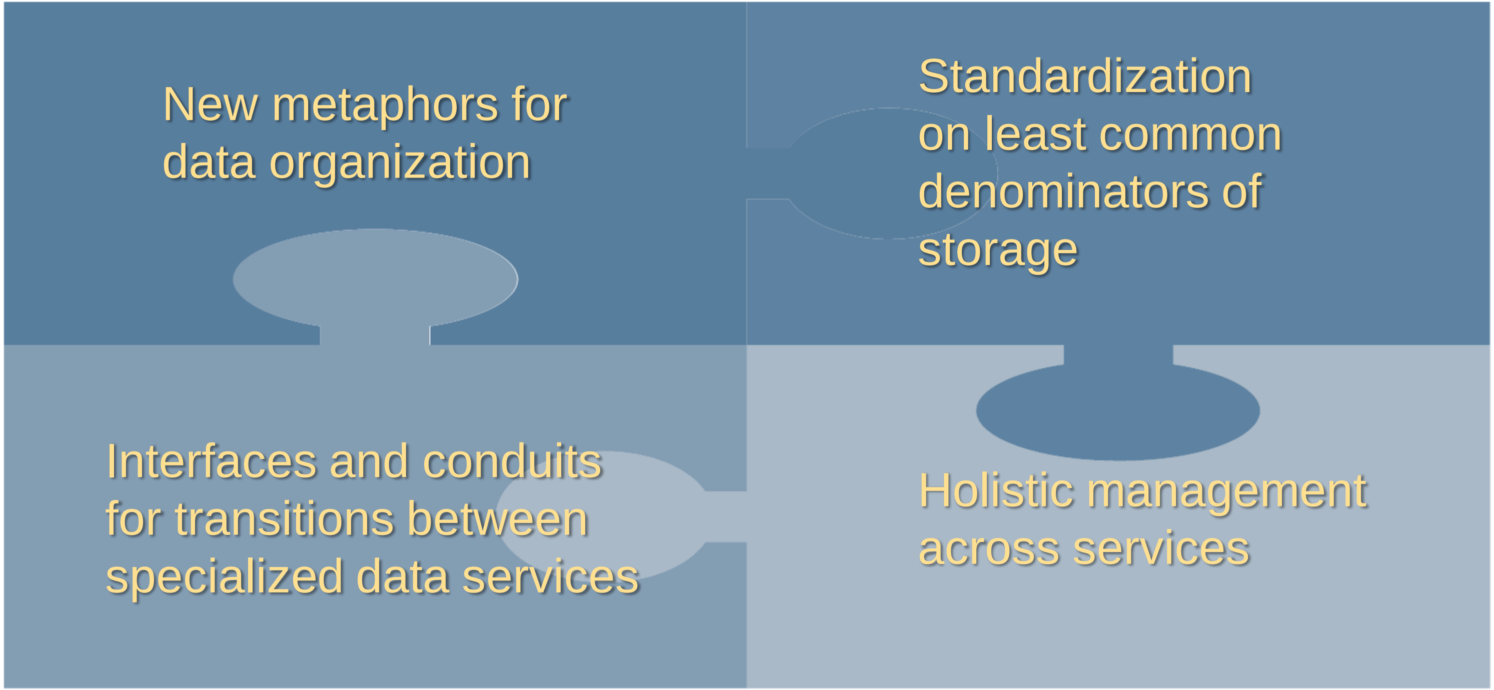
- **Amazon Timestream**

IF THEY CAN DO IT, WHY NOT US?

Why doesn't HPC have a similar storage ecosystem?
Or better yet, why can't you Bring Your Own File System?

- Risk of administrative/procurement cost explosion
  - We can't support N separately administrated silos.
- Data stewardship
  - How do you make sure that mission critical data is safe, persistent, available, and **portable** when it's scattered across devices and services?
- Infrastructure
  - Amazon has built a home-grown infrastructure to support different models.
  - Storage vendors would be reluctant to do so on their own.
- Occasional philosophical tangents
  - What storage system is the "best"? Is POSIX dead?

WHAT WOULD IT TAKE TO ENABLE BYOFS?



New metaphors for data organization

Standardization on least common denominators of storage

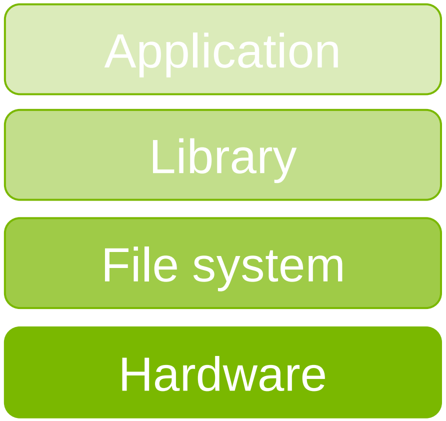
Interfaces and conduits for transitions between specialized data services

Holistic management across services

Argonne National Laboratory is a U.S. Department of Energy laboratory managed by UChicago Argonne, LLC.

A NEW METAPHOR FOR HPC STORAGE

The traditional HPC stack



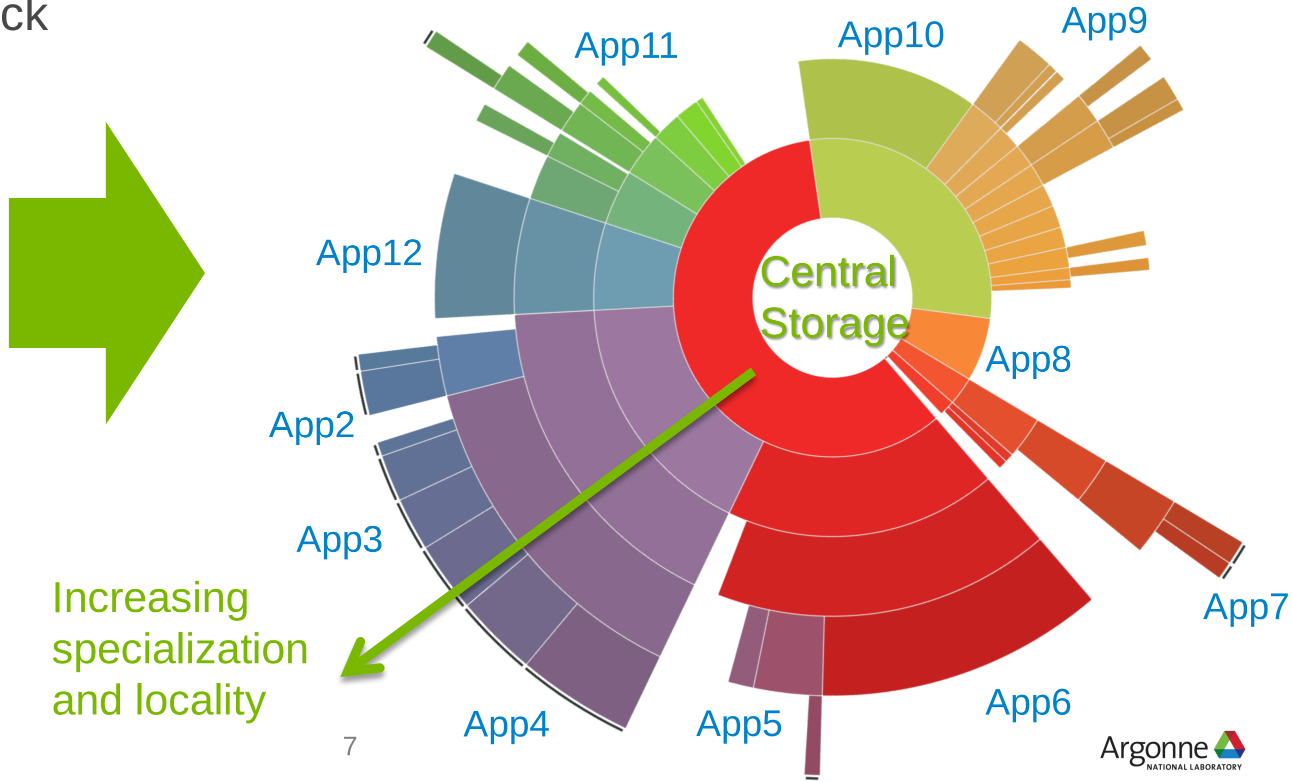
Application

Library

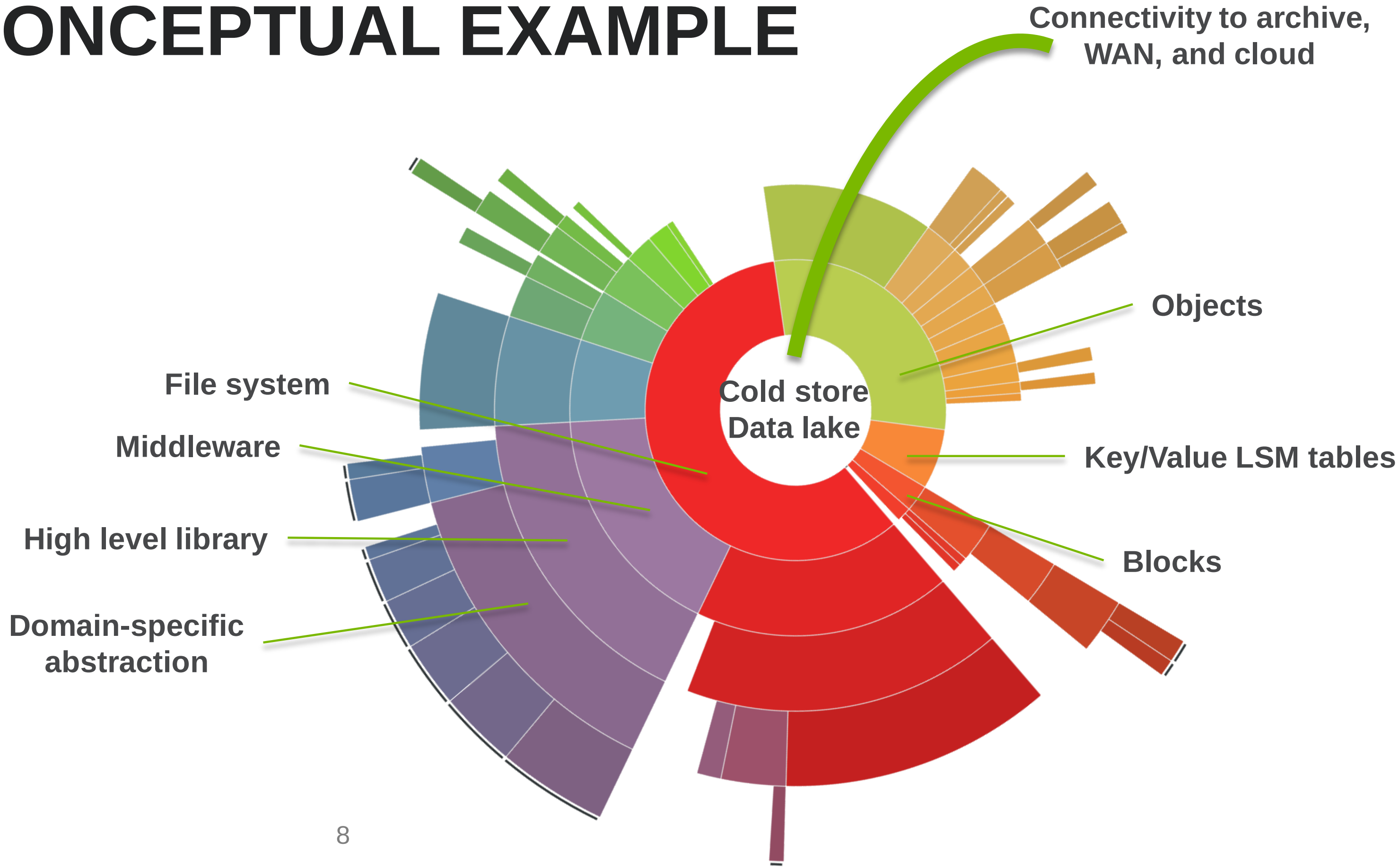
File system

Hardware

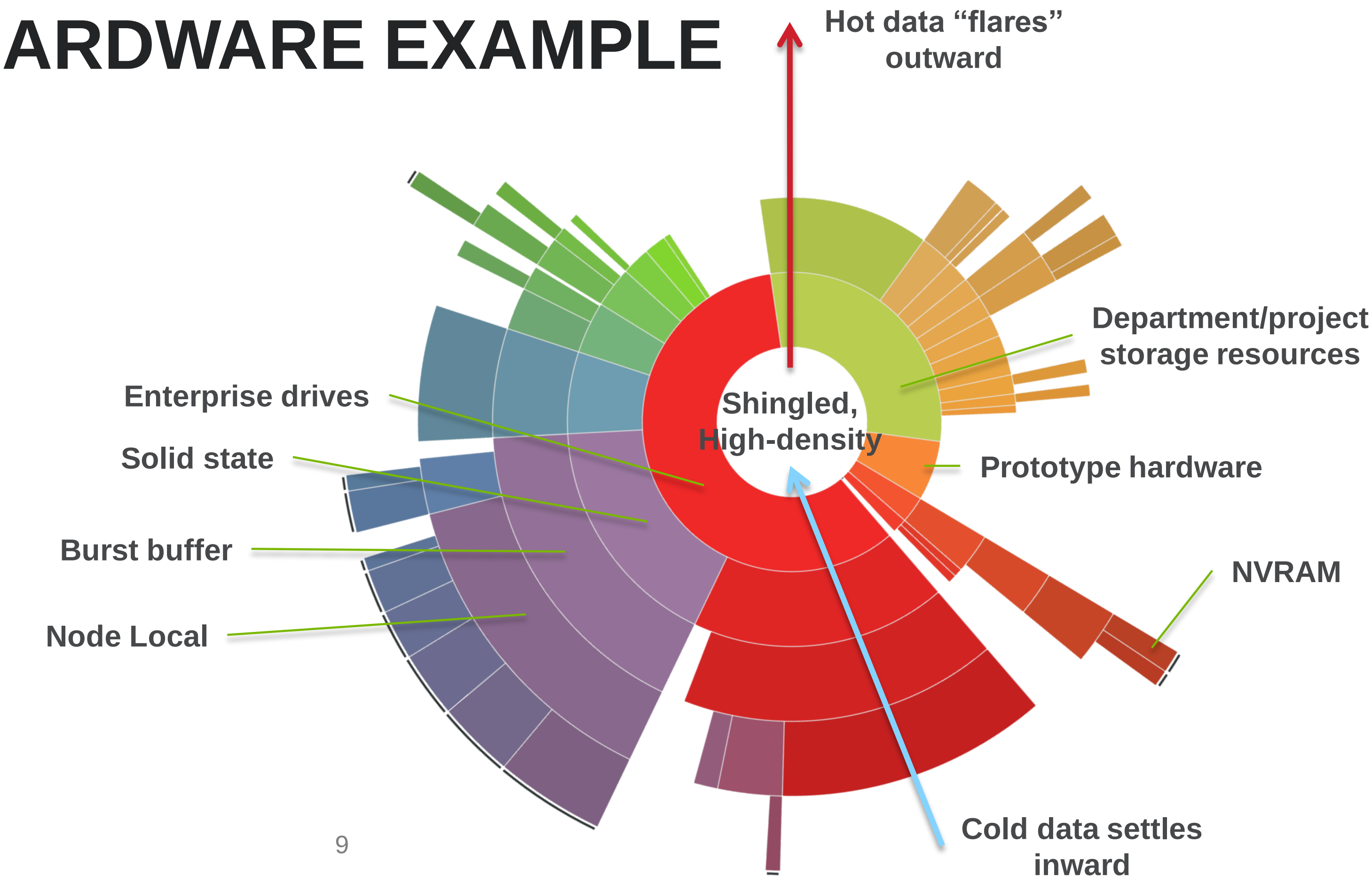
Many different HPC “stacks” radiating outward from a commonly managed central core



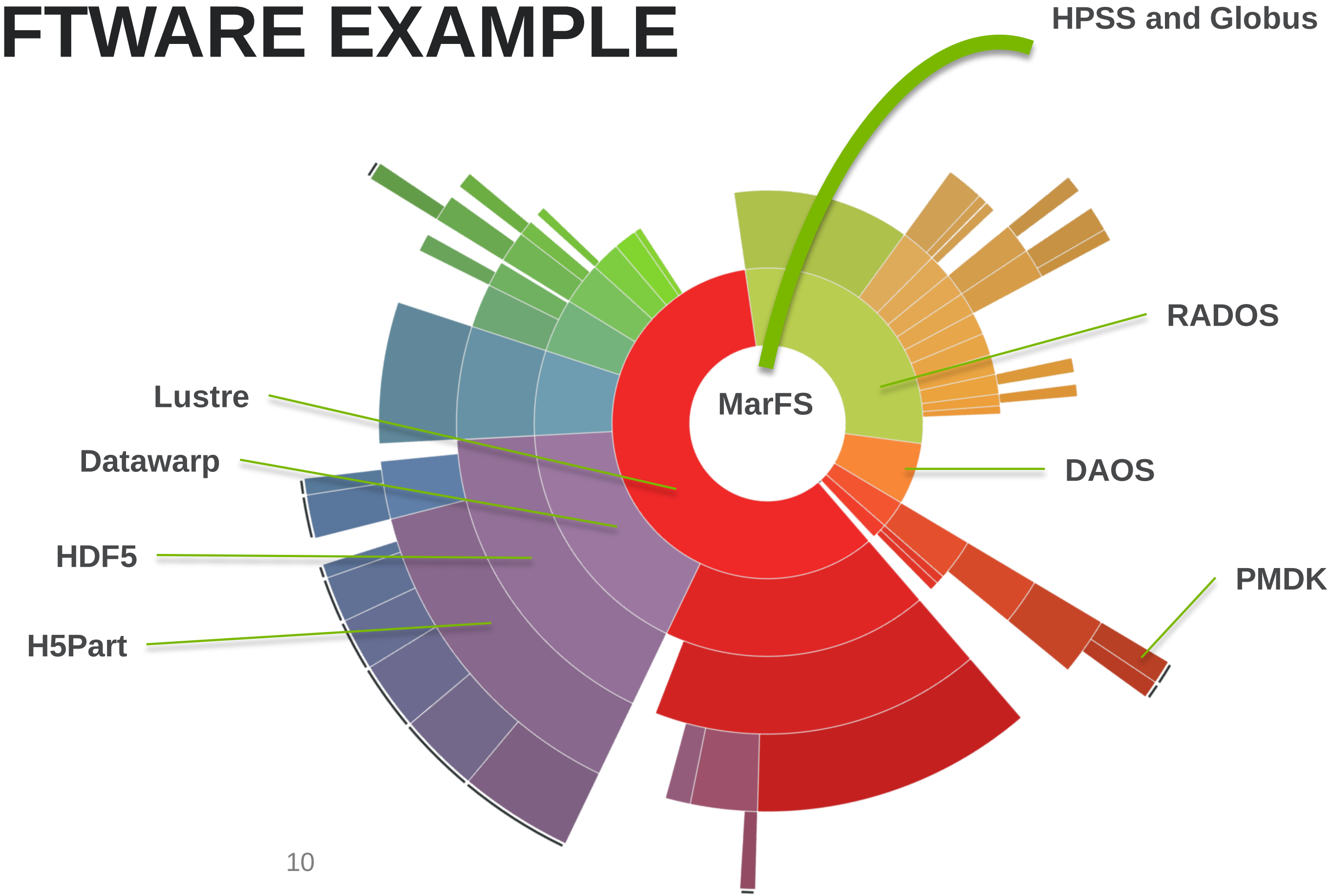
CONCEPTUAL EXAMPLE



HARDWARE EXAMPLE



SOFTWARE EXAMPLE

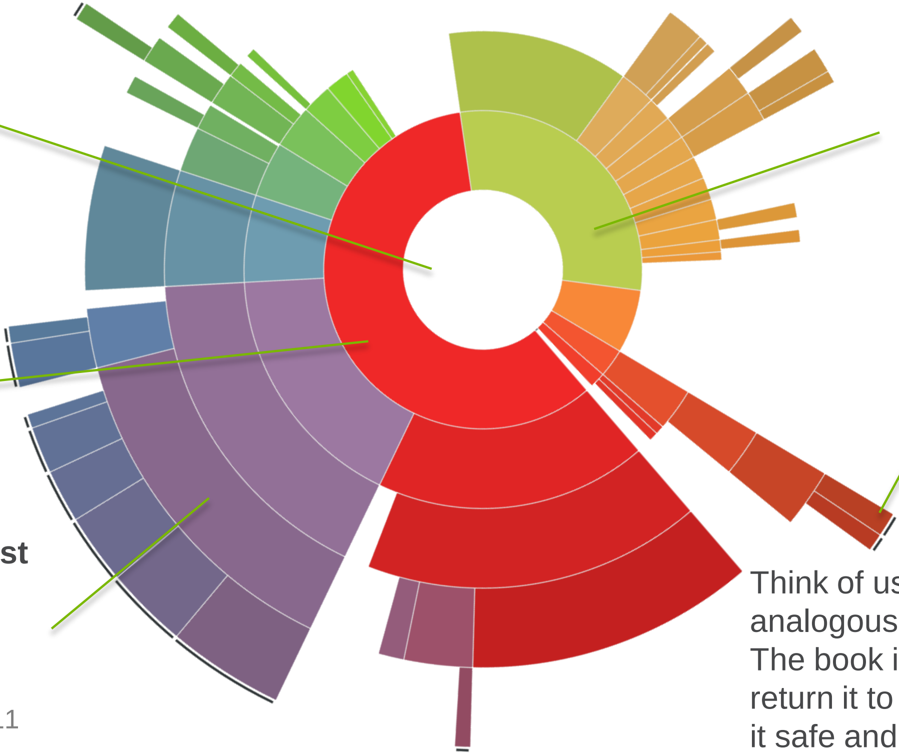


POLICY EXAMPLE

Only data that percolates down to the central store is directly managed by facility, guaranteed resilient, available to WAN, etc.

Layer 1 services provided by facility, but dynamically provisioned into non-global pools.

BYOFS: domain services must conform to transit/conduit APIs to participate in the ecosystem. Little facility intervention at this level.



Some departments may fund and provision compatible resources for QOS or unusual connectivity.

Some teams may be extraordinarily specialized.

Think of using data from the central core as analogous to checking out a library book. The book is your responsibility until you return it to the library. Then the library keeps it safe and makes it available to others.

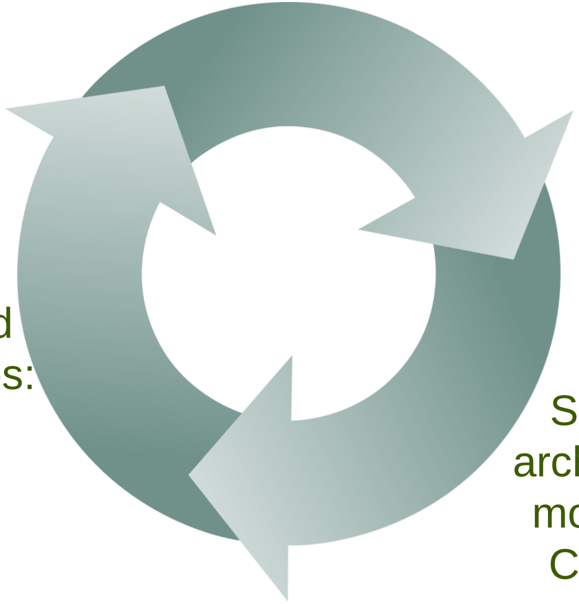
CHALLENGES TO OVERCOME

- What are the interfaces and/or conduits between layers?
  - This will be the “must be this tall to ride” requirement for new components to participate in this model.
- Provisioning: How big are the outer layers, who gets them, and for how long?
  - Are some granted with project allocation?
  - Are some granted when the job is scheduled?
  - Can the provisioning look transparent to users?
    - How about if it is oversubscribed and staged out when not in use?
- Policies: how make people use resources responsibly when they aren’t being billed per GiB?
  - Come to think of it, we have that challenge already. This isn’t a new problem.

COMPLEMENTARY DATA RESEARCH AT ANL

Some building blocks

- Mochi:
  - http://www.mcs.anl.gov/research/projects/mochi/
- TOKIO (and Darshan):
  - https://www.nersc.gov/research-and-development/tokio/
  - http://www.mcs.anl.gov/research/projects/darshan/
- CODES:
  - http://www.mcs.anl.gov/research/projects/codes/



Holistic observation: TOKIO / Darshan

Specialized data services: Mochi

Storage architecture modeling: CODES

THANK YOU!

THIS WORK WAS SUPPORTED BY THE U.S. DEPARTMENT OF ENERGY, OFFICE OF SCIENCE, ADVANCED SCIENTIFIC COMPUTING RESEARCH, UNDER CONTRACT DE-AC02-06CH11357.