Policy Based Data Management or Moving Computation to the Data

Reagan W. Moore

Arcot Rajasekar

Mike Wan

{moore,sekar,mwan}@diceresearch.org

http://irods.diceresearch.org

D·I·C·E

i·R·O·D·S

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UCSD

NSF

THE NATIONAL ARCHIVES ARCHIVES.GOV

SRB

Observations

- Data processing pipelines to support data mining
  - Data-driven science based on data mining
  - Detect significant events
  - Generate statistics by varying input conditions
  - Apply data processing pipelines to generate standard products
- Digital libraries to support publication within a discipline
  - Provide services for use of the collection
- Preservation as reference collections
  - Digital holdings on which future research is based

- Multiple types of data management environments

Digital Library

- Texas Digital Library
- French National Library

Data Processing Pipeline

- Ocean Observatories Initiative
- Large Synoptic Survey Telescope

Data Grid

- Teragrid
- Temporal Dynamics of Learning Center
- Australian Research Collaboration Service

Preservation Environment

- NARA Transcontinental Persistent Archive Prototype
- Taiwan National Archive
- Carolina Digital Repository

D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV SRB

Observations (cont.)

- Observe that many projects are generating massive data collections
  - Observational data (astronomy, climate change, oceanography)
  - Experimental data (high energy physics, biology)
  - Simulation output (high energy physics, seismology, earth systems, cosmology)
- Data are widely distributed
  - Sources, storage systems, analysis systems, users
- Scale is now hundred petabytes, hundreds of millions of files

Cloud Storage

amazon web services™

MOSSO the hosting cloud

GO GRID

Institutional Repositories

Carolina Digital Repository

Texas Digital Library

Federal Repositories

National Climatic Data Center

National Optical Astronomy Observatory

D·I·C·E

i·R·O·D·S

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

UCSD

NSF

THE NATIONAL ARCHIVES ARCHIVES.GOV

Questions

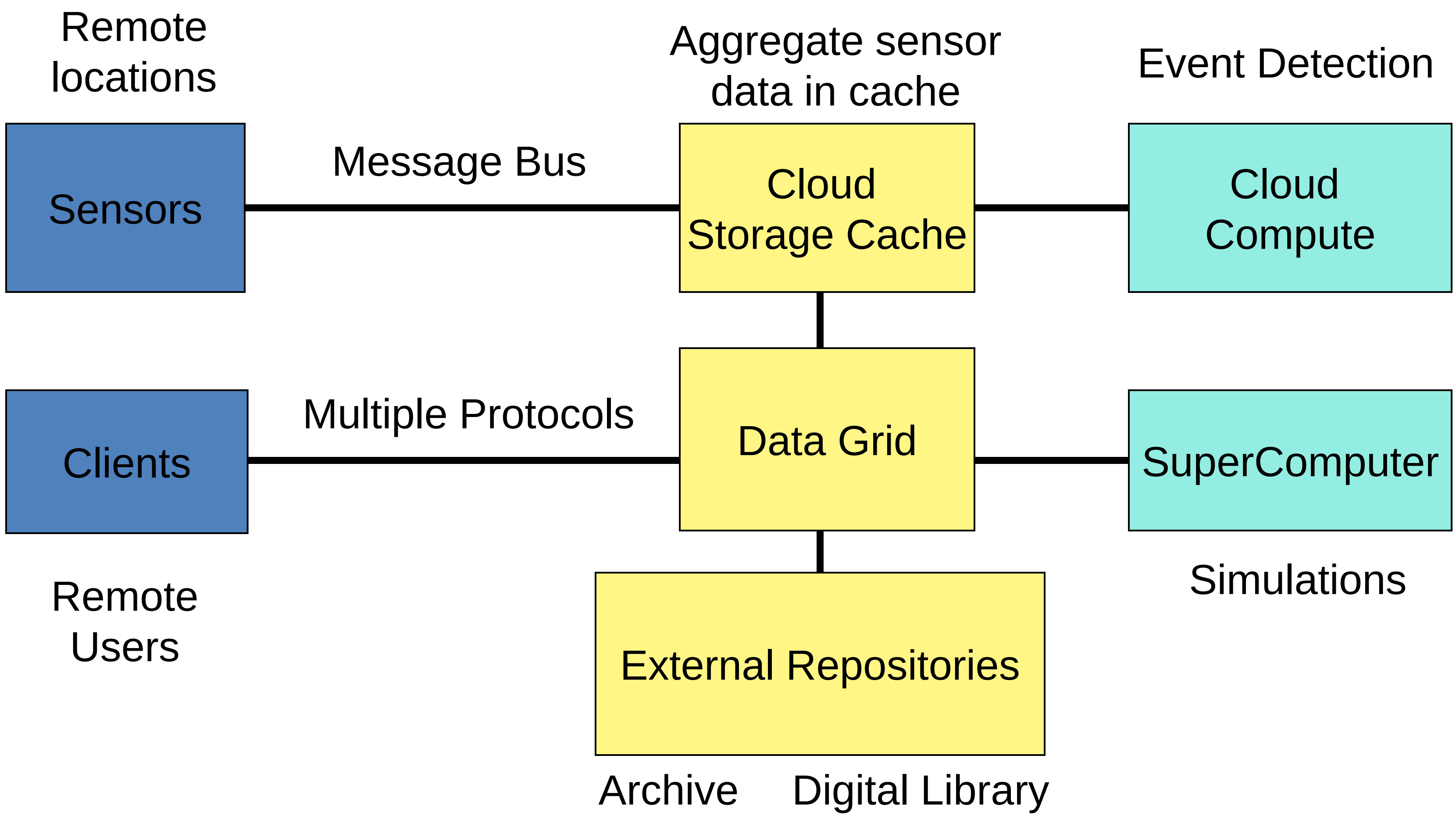
- Can these multiple environments be integrated?
- Where should data be stored within these systems?
- Where should the data be analyzed?

- Data grids: support remote processing of data

Distributed Workflows

- When are data processed at the remote storage location?
  - Low complexity operations
- When are data processed at a supercomputer?
  - High complexity operations
- When are data processed at the display?
  - Interactive presentation manipulation

Ocean Observatories Initiative



THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

“Ohm’s” Law for Computer Science

- Relationship between
  - Computational complexity (operations per byte)
  - Execution rate
  - Data access bandwidth

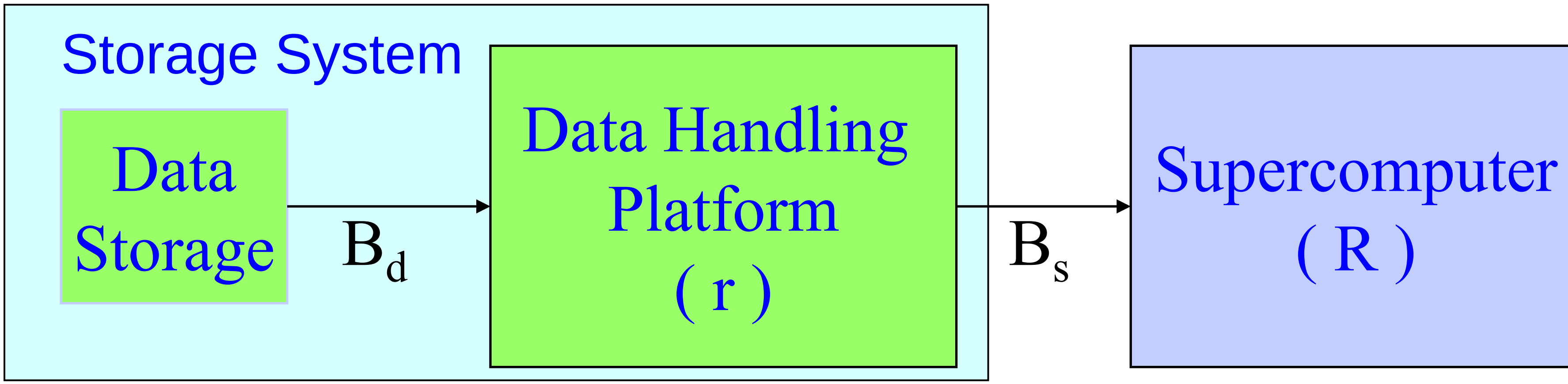
$$\eta = R / B$$

Complexity = Execution Rate / Bandwidth

for a balanced application

Data Distribution Thought Experiment

Reduce size of data from S bytes to s bytes and analyze



Execution rates are	$r < R$
Bandwidths linking systems are	$B_d > B_s$
Operations per byte for analysis is	η_s
Operations per byte for data transfer is	η_t

Should the data reduction be done before transmission?

Distributing Services

Compare times for analyzing data with size reduction from S to s

Data Handling Platform				Supercomputer
Read Data	Reduce Data	Transmit Data	Network	Receive Data
S / B_d	$\eta_s S / r$	$\eta_t s / r$	s / B_s	$\eta_t s / R$

Data Handling Platform			Supercomputer	
Read Data	Transmit Data	Network	Receive Data	Reduce Data
S / B_d	$\eta_t S / r$	S / B_s	$\eta_t S / R$	$\eta_s S / R$

Comparison of Time

Processing at archive

$$T(\text{Archive}) = S/\ B_d + \eta_s\ S/r + \eta_t\ s/r + s/\ B_s + \eta_t\ s/R$$

Processing at supercomputer

$$T(\text{Super}) = S/\ B_d + \eta_t\ S/r + S/\ B_s + \eta_t\ S/R + \eta_s\ S/R$$

Selecting Analysis Location

Have algebraic equation with eight independent variables.
Faster to move the data if:

T (Super) < T (Archive)

$$S/B_d + \eta_t S/r + S/B_s + \eta_t S/R + \eta_s S/R$$

$$< S/B_d + \eta_s S/r + \eta_t s/r + s/B_s + \eta_t s/R$$

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Scaling Parameters

Data size reduction ratio	s/S
Execution slow down ratio	r/R
Problem complexity	η_t / η_s
Communication/Execution	$r/(\eta_t B_s)$

Note (r/η_t) is the number of bytes/sec that can be processed.

When $r/(\eta_t B_s) = 1$, the data processing rate is the same as the data transmission rate.

Optimal designs have $r/(\eta_t B_s) = 1$

Bandwidth Optimization

Is moving all of the data faster, T(Super) < T(Archive), if the network is sufficiently fast?

$$B_s > (r / \eta_s) (1 - s/S) / [1 - r/R - (\eta_t / \eta_s) (1 + r/R) (1 - s/S)]$$

Note the denominator changes sign when

$$\eta_s < \eta_t (1 + r/R) / [(1 - r/R) (1 - s/S)]$$

Even with an infinitely fast network, it is better to do the processing at the archive if the complexity is too small.

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Execution Rate Optimization

Is moving all of the data faster, T(Super) < T(Archive), if the supercomputer is sufficiently fast?

$R > r\, [1 + (\eta_t / \eta_s)\, (1 - s/S)] / [1 - (\eta_t / \eta_s)\, (1 - s/S)\, (1 + r/(\eta_t\, B_s)]$

Note the denominator changes sign when
$\eta_s < \eta_t\, (1 - s/S)\, [1 + r/(\eta_t\, B_s)]$

Even with an infinitely fast supercomputer, it is better to process at the archive if the complexity is too small.

Data Reduction Optimization

Is processing at the archive faster, T(Super) > T(Archive), if the data reduction is large enough?

$s < S \{1 - (\eta_s / \eta_t)(1 - r/R) / [1 + r/R + r/(\eta_t B_s)]\}$

Note criteria changes sign when
$\eta_s > \eta_t [1 + r/R + r/(\eta_t B_s)] / (1 - r/R)$

When the complexity is sufficiently large, it is faster to process on the supercomputer even when data can be reduced to one bit.

Complexity Analysis

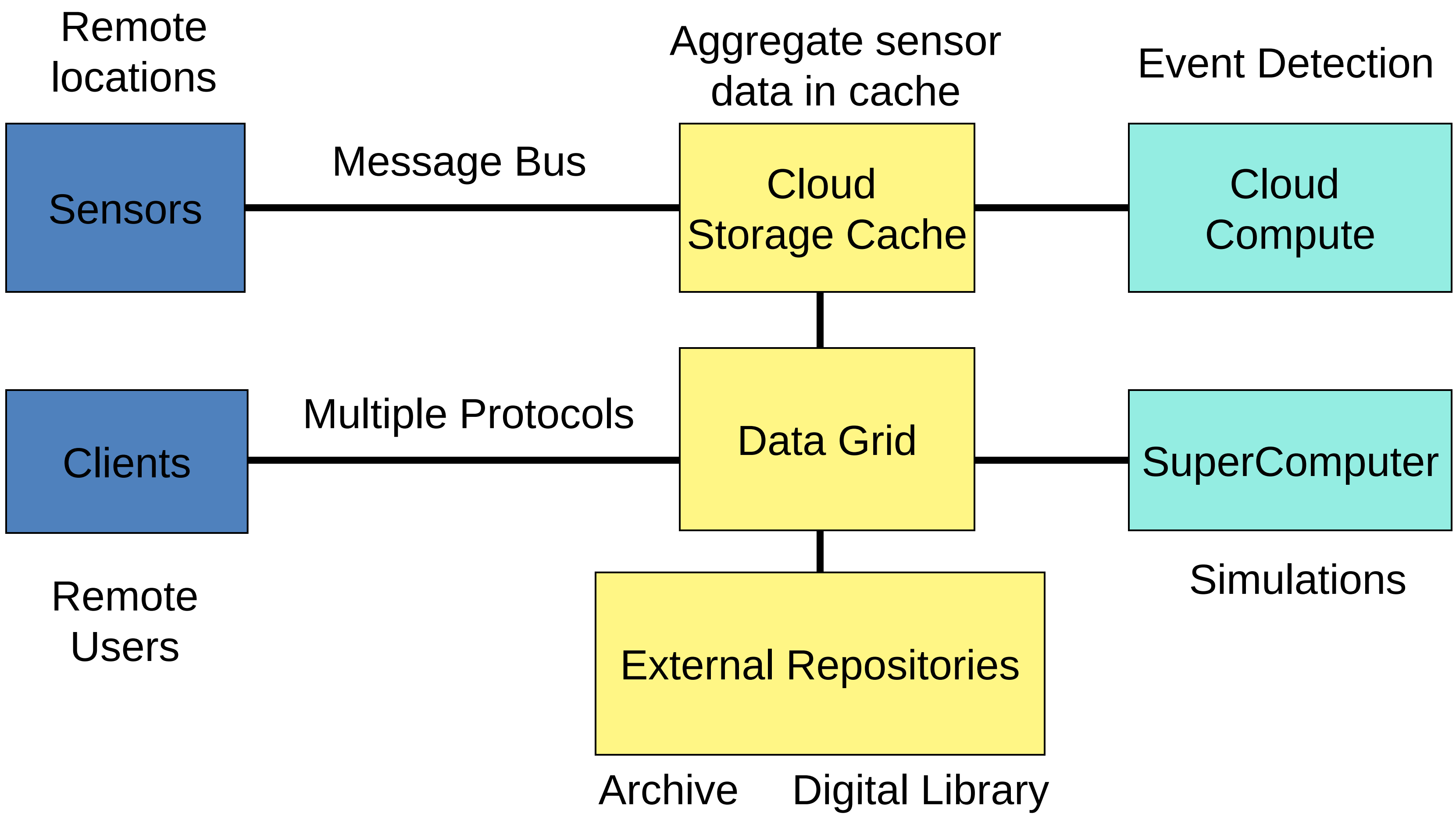
Moving all of the data is faster, T(Super) < T(Archive) if the complexity is sufficiently high!

$$\eta_s > \eta_t (1\text{-}s/S) [1 + r/R + r/(\eta_t B_s)] / (1\text{-}r/R)$$

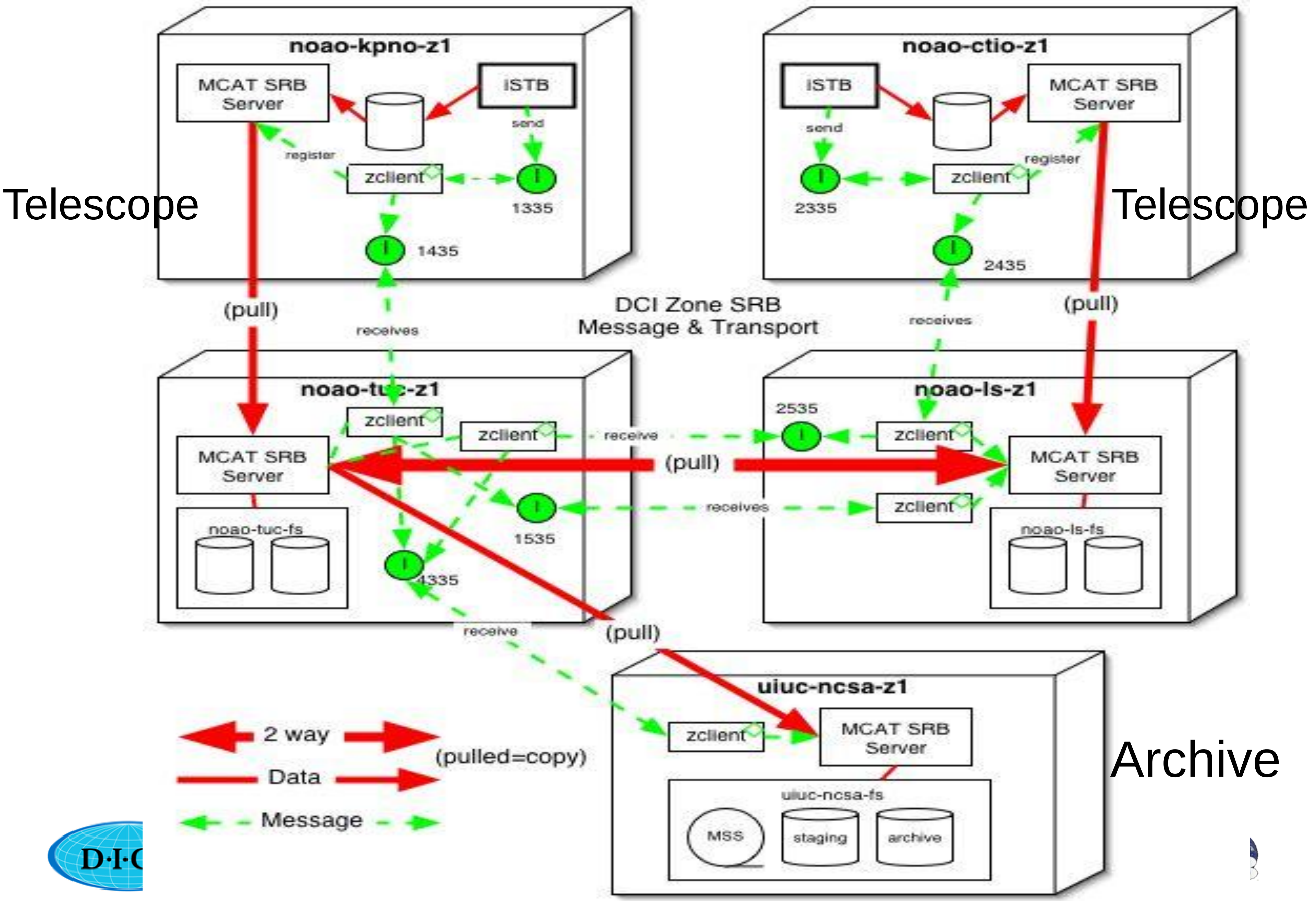
Note, as the execution ratio approaches 1, the required complexity becomes infinite

Also, as the amount of data reduction goes to zero, the required complexity goes to zero.

Ocean Observatories Initiative



NOAO Zone Architecture



Policy-based Data Environments

- *Purpose* - reason a collection is assembled
- *Properties* - attributes needed to ensure the **purpose**
- *Policies* - control for ensuring maintenance of **properties**
- *Procedures* - functions that implement the **policies**
- *State information* - results of applying the **procedures**
- *Assessment criteria* - validation that **state information** conforms to the desired **purpose**
- *Federation* - controlled sharing of **logical name spaces**

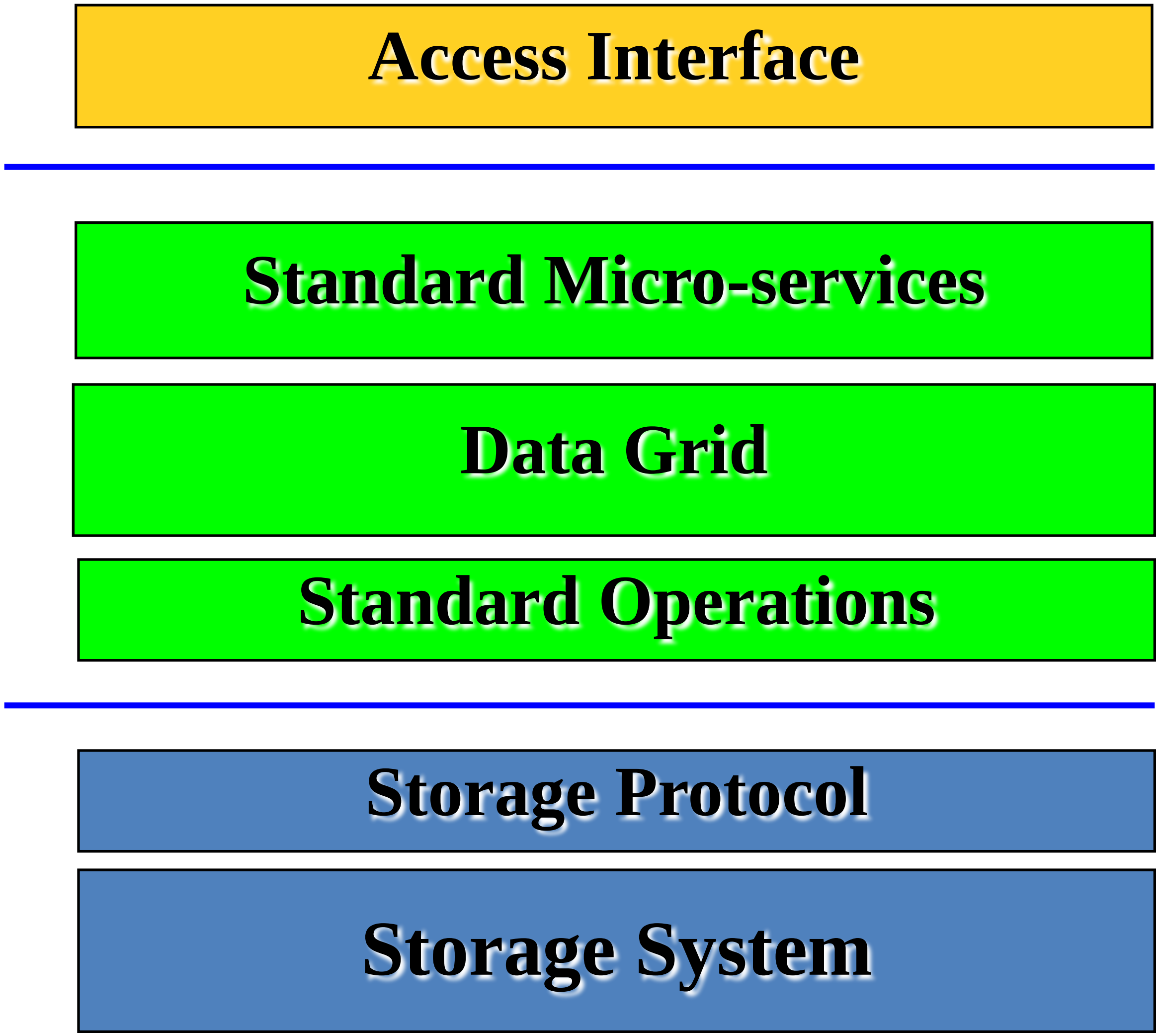
These are the necessary elements for a sustainable collection

D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV SRB

iRODS - Policy-based Data Management

- Turn policies into computer actionable rules
- Compose rules by chaining standard operations
  - Standard operations (micro-services) executed at the remote storage location
- Manage state information as attributes on namespaces:
  - Files / collections /users / resources / rules
- Validate assessment criteria
  - Queries on state information, parsing of audit trails
- Automate administrative functions
  - Minimize labor costs

Data Virtualization



Map from the actions requested by the access method to a standard set of micro-services.

The standard micro-services are mapped to standard operations.

The standard operations are mapped to the protocol supported by the storage system

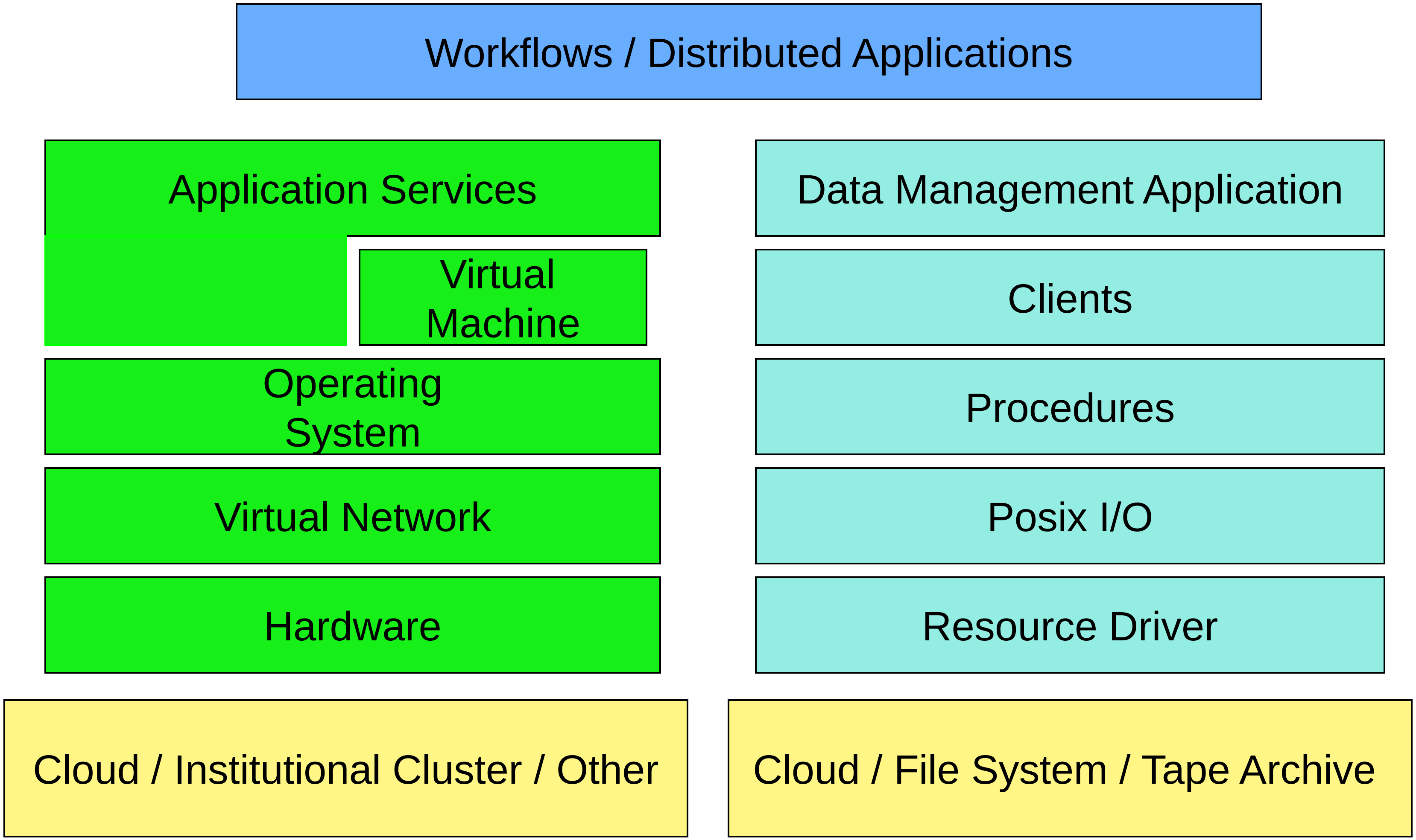
Data Grid Clients

API	Client	Developer
Browser		
	DCAPE	UNC
	iExplore	DICE-Bing Zhu
	JUX	IN2P3
	Peta Web browser	PetaShare
Digital Library		
	Akubra/iRODS	DICE
	Dspace	MIT
	Fedora on Fuse	IN2P3
	Fedora/iRODS module	DICE
	Islandora	DICE
File System		
	Davis - Webdav	ARCS
	Dropbox / iDrop	DICE-Mike Conway
	FUSE	IN2P3, DICE,
	FUSE optimization	PetaShare
	OpenDAP	ARCS
	PetaFS (Fuse)	Petashare - LSU
	Petashell (Parrot)	PetaShare
Grid		
	GridFTP - Griffin	ARCS
	Jsaga	IN2P3
	Parrot	Notre Dame-Doug Thain
	Saga	KEK

API	Client	Developer
I/O Libraries		
	PHP - DICE	DICE-Bing Zhu
	C API	DICE-Mike Wan
	C I/O library	DICE-Wayne Schroeder
	Jargon	DICE-Mike Conway
	Pyrods - Python	SHAMAN-Jerome Fusillier
Portal		
	EnginFrame	NICE / RENCI
Tools		
	Archive tools-NOAO	NOAO
	Big Board visualization	RENCI
	File-format-identifier	GA Tech
	icommands	DICE
	Pcommands	PetaShare
	Resource Monitoring	IN2P3
	Sync-package	Academica Sinica
	URSpace	Teldap - Academica Sinica
Web Service		
	VOSpace	NVOA
	Shibboleth	King's College
Workflows		
	Kepler	DICE
	Stork	LSU
	Taverna	RENCI

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Virtualization Stacks



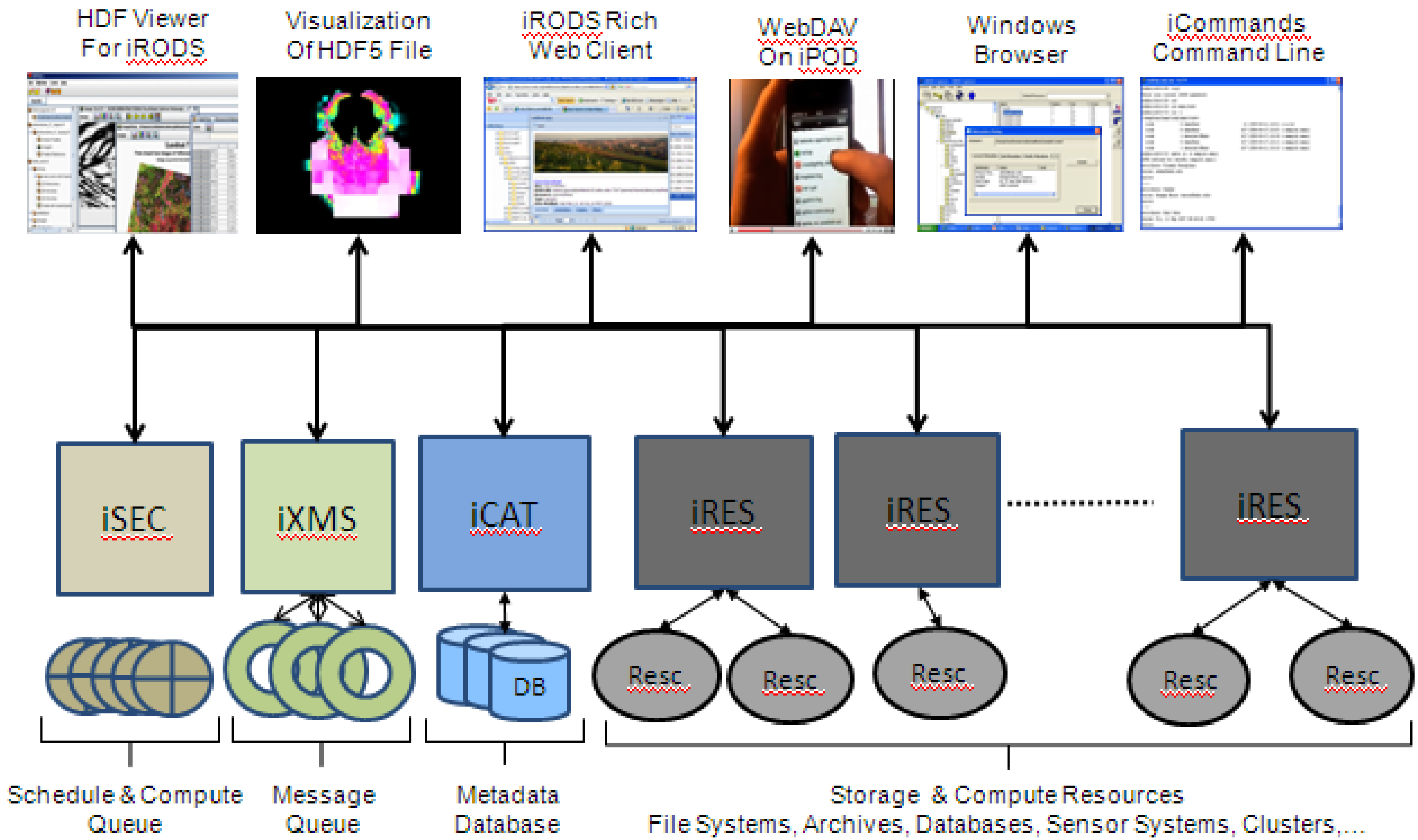
Storage Cost Scaling
(as media capacity increases)

- For large scale systems:
  - Capital investment (33%)
    - Tape robot, tape drives — Scales with Technology
  - Media (33%)
    - Tape cartridges — Scales with Technology
  - Operations (33%)
    - Software licenses — Scales with Technology
    - Facilities — Scales with Technology
    - Administration — Need automation

Infrastructure Development Costs

- Storage Resource Broker middleware development
  - 300,000 lines of code
  - Six year development / ten year deployment
  - 10-15 professional software engineers
- Total cost ~ $15,000,000
  - $17 / line for design, development, testing, documentation, bug fixes
  - $14 / line for interoperability (clients)
  - $12 / line for application use support
  - $7 / line for management / administration
  - Total cost ~ $50 / line
- Development funded by:
  - NSF / NARA / DARPA / DoE / NASA / NIH / IMLS / NHPRC / LoC / DoD
  - More than 20 funded projects to sustain development
  - International collaborations on use, development, bug fixes, support

iRODS Distributed Data Management



D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV

Goal - Generic Infrastructure

- Manage all stages of the data life cycle
  - Data organization
  - Data processing pipelines
  - Collection creation
  - Data sharing
  - Data publication
  - Data preservation
- Create reference collection against which future information and knowledge is compared
  - Each stage uses similar storage, arrangement, description, and access mechanisms

Data Life Cycle

Each data life cycle stage re-purposes the original collection

Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	Shared	Analyzed	Published	Preserved	Sustained
Local Policy	Distribution Policy	Service Policy	Description Policy	Representation Policy	Re-purposing Policy

Stages correspond to addition of new policies for a broader community
Virtualize the stages of the data life cycle through policy evolution

Demonstration

- Data grid in North Carolina at RENCI
- Icommands user interface (file manipulation)
- System state information
- Rule base controlling the data grid (policies)
- Composition of rules from micro-services
- Interactive execution of server-side workflows

iRODS is a "coordinated NSF/OCI-Nat'l Archives research activity" under the auspices of the President's NITRD Program and is identified as among the priorities underlying the President's 2009 Budget Supplement in the area of Human and Computer Interaction Information Management technology research.

Reagan W. Moore

rwmoore@renci.org

http://irods.diceresearch.org

NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"

NSF SDCI-0721400 "Data Grids for Community Driven Applications"

D·I·C·E i·R·O·D·S THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL UCSD NSF THE NATIONAL ARCHIVES ARCHIVES.GOV SRB