Integrating Data Life Cycle into Mission Life Cycle

Arcot Rajasekar

rajasekar@unc.edu

sekar@diceresearch.org

Technology of Interest

Provide an end-to-end capability for

Exa-scale data orchestration

From creation & sensing (immediate use)

to analysis & assimilation (modeling)

to discovery & sharing (distribution)

to archiving & disposition (long-term)

From Design to mission completion to cross-over

Need for integrated data and metadata system

Development of a data-intensive cyber environment - integrating flight and ground capabilities - should be a focus that can help all areas of mission life-cycle

Need for this technology

- Sample Points of Interest from Roadmap
  - Unification of aerospace engineering, remote science, current space operations, and future human exploration missions
  - Digital Twin
  - Adaptable data management and analytics
  - Energy-aware systems (data placement strategies)
  - Integrated H/W and S/W modeling
  - Distributed Simulation
  - Integrated System Lifecycle Simulation
  - Deal with multi-decadal data and multi-domain data
  - Forensics and debugging
  - High fidelity configurable simulation
  - Intelligent Data Undestanding (IDU)
    - defines an impressive set of needs
  - Collaborative Science and Engineering

Current Roadmap

- Showcases the need for data-intensive capability at various levels in TA11
- Provides limited guidance as to how to pull and push this technology
- Information Processing Roadmap is very impressive but needs a corresponding ‘evolutionary’ data orchestration roadmap
- Game-changing ‘State of art’ is available in terms of
  - Policy-oriented data life-cycle management
  - Data virtualization technologies (*-agnostic)
  - Service-oriented data operations (flexible and configurable)
    - Orchestrate semantically well-defined services
  - Distributed cloud storage and computing (third party – on demand)

Policy-based Data Environments

- *Purpose* - reason a dataset is assembled
- *Properties* - attributes needed to ensure the **purpose**
- *Policies* - controls for enforcing desired **properties,**
- **mapped to computer actionable rules**
- *Procedures* - functions that implement the **policies**
- **mapped to computer actionable workflows**
- *State information* - results of applying the **procedures**
- **mapped to system metadata**
- *Assessment criteria* - validation that **state information** conforms to the desired **purpose**
- **mapped to periodically executed policies**

Policies for Data Ops

Instead of *managing bytes* and file

manage the policies and

let the *policy engine* manage the bytes and files

- Data Placement - co-location, distribution, disposition
- Semantic linking - descriptive metadata
- Fault tolerance - replication, synchronization, versioning
- Lifecycle – create, cache, share, dessiminate,arcchive
- Privacy -authentication, authorization, audit
- Security - Assessment criteria, validation, integrity checks
- Assimilation - derived data product generation
- Federation - seamless spanning across distributed, autonomous data collections
- Stream processing – on ingest, on access, in store

NASA's Effort & Time Horizons

- Currently, NASA Center for Computational Sciences is applying policy-based data management to simulation frameworks
  - Application to share data with NCDC/NOAA under consideration
- Lot of requirement for cloud storage
- Time Horizons – in conjunction with IP roadmap
  - 5-10 years – ground data federation
  - 10-20 years – flight data federation
  - 20-30 years – seamless data federation across ground-flight boundaries
- Players: NASA, academic & federal agencies (NSF, NOAA, NARA, DOE, …)

Barriers, Risks, PayOffs

Comprehensive data orchestration strategy

- Paradigm shift – data-oriented, policy-oriented, outcome-oriented

  (capture behavior in terms of data outcomes)
- Risk – Can policies and services capture all the intricacies of NASA mission requirements
- PayOffs – Seamless/extensible system, tolerance, integrated semantics, verfiable design.

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.

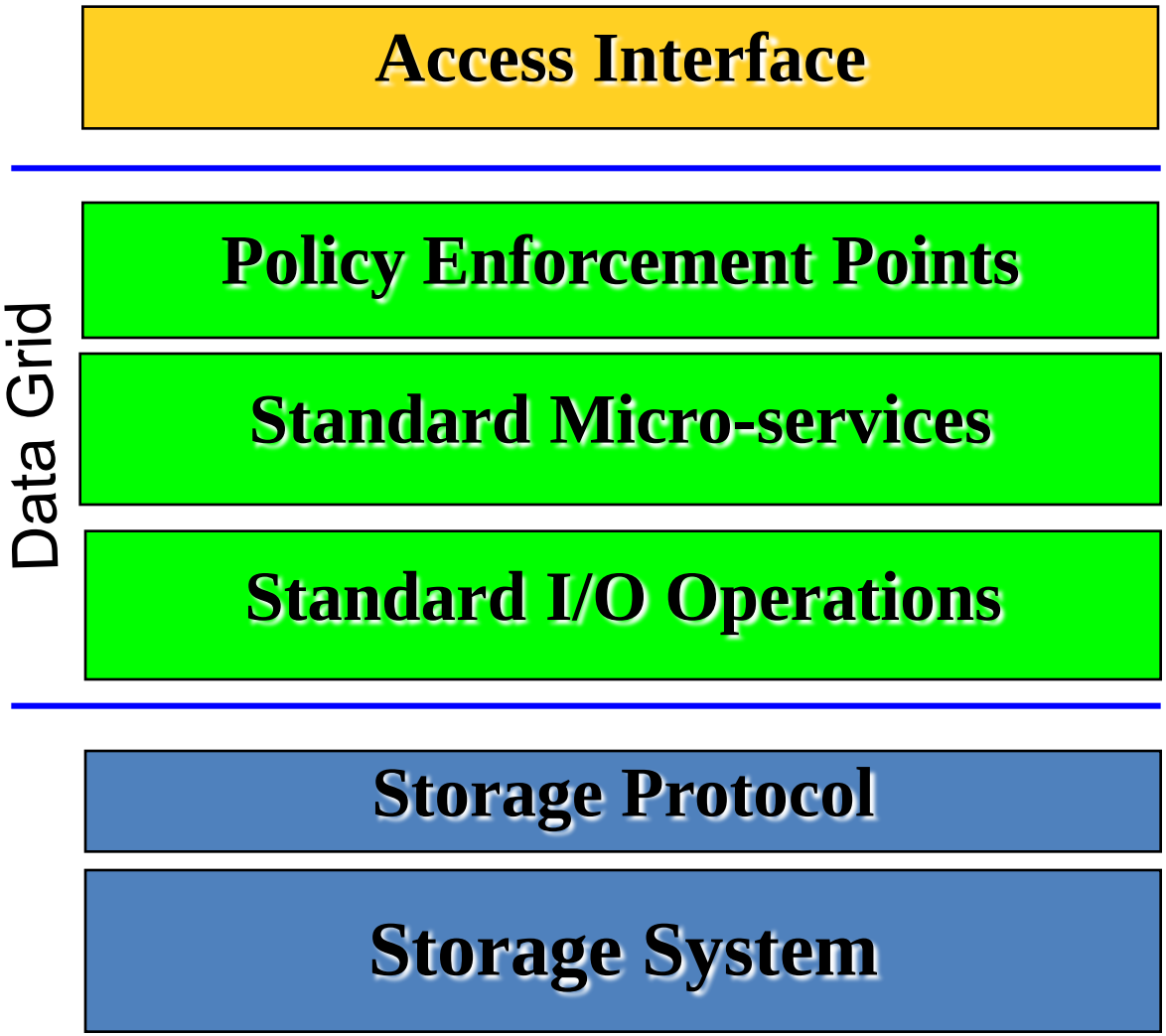
Arcot Rajasekar

rajasekar@unc.edu

http://irods.diceresearch.org

NSF OCI-0848296 "NARA Transcontinental Persistent Archives Prototype"
NSF SDCI-0721400 "Data Grids for Community Driven Applications"

Data Virtualization



Map from the actions requested by the client to multiple policy enforcement points.

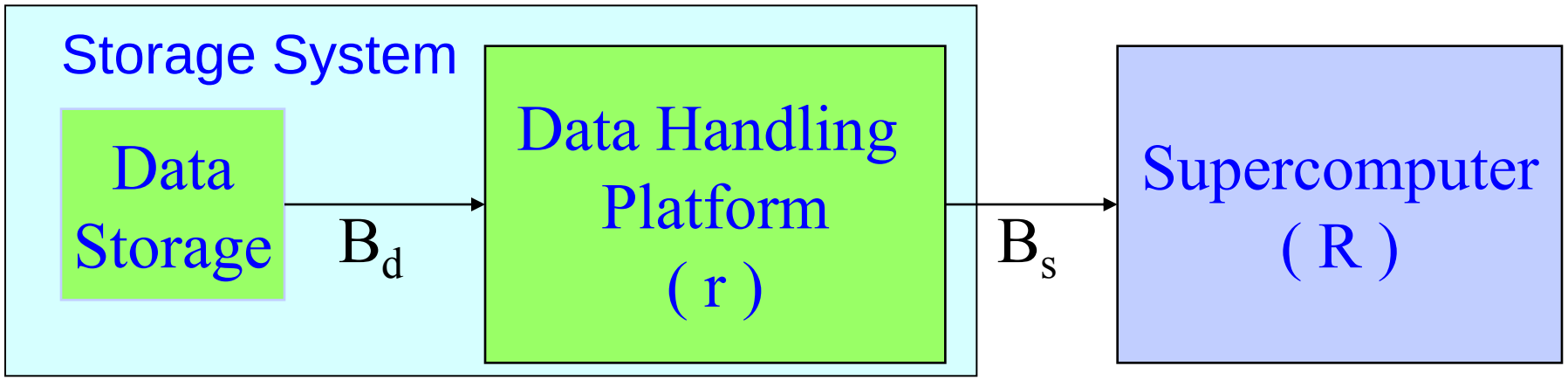
Map from policy to standard micro-services.

Map from micro-services to standard Posix I/O operations.

Map standard I/O operations to the protocol supported by the storage system

Data Distribution Thought Experiment

Reduce size of data from S bytes to s bytes and then 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?

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.

For sufficiently low complexity, it is faster to do the computation at the storage location

Applications

- Data grids
  - Astronomy – NOAO, CyberSKA, LSST
  - High Energy Physics – BaBar. KEK
  - Earth Systems – NASA MODIS data set
- Institutional repositories
  - Carolina Digital Repository
- Libraries
  - Texas Digital Libraries
  - Seismology - Southern California Earthquake Center
- Archives
  - Ocean Observatories Initiative

Micro-Services

- Small, well-defined "compiled" functions
  - Does a particular task (can be more than one way of achieving the task)
  - Semantics defined in terms of "side-effects"
- Chained together and interpreted at run-time
  - To achieve composite goals (based on side-effects) – easy verification
  - Invoked on conditions, events, triggers, user commands, or periodic
- Multiple modes of Interactions between micro-services
  - InBand: Parameters, Shared Whiteboard (memory-level interactions)
  - OutOfBand: Metadata Catalog, Message System (external services)
- Provide:
  - Standard operations/functions as needed by the domain
  - Invocation of external applications & Interaction with web services
  - Launch jobs in cloud & workflows systems
  - Remote, concurrent or delayed execution control - between micro-servers.