THE UNIVERSITY OF CHICAGO

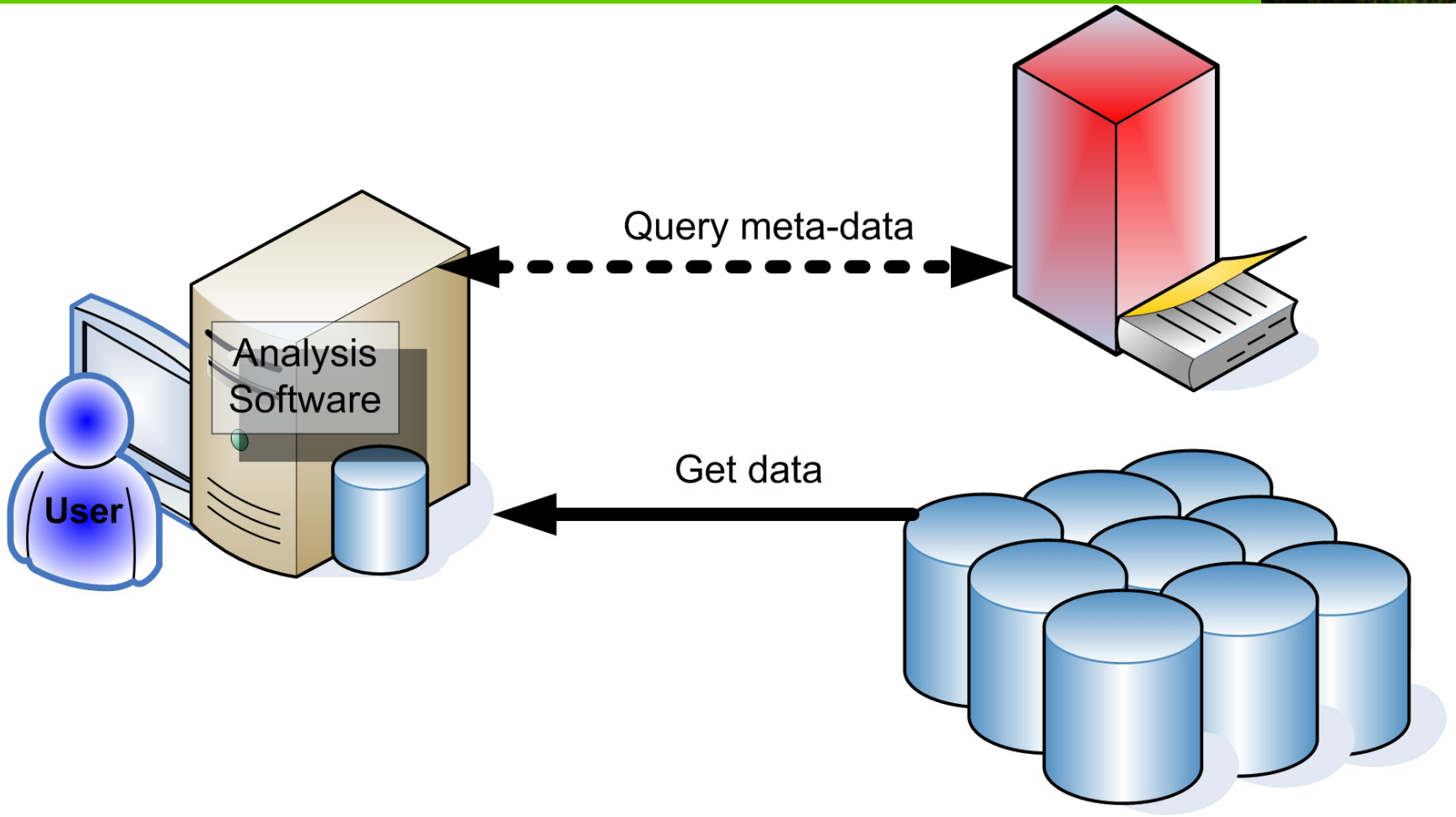
ARGONNE NATIONAL LABORATORY

Storage and Compute Resource Management via DYRE, 3DcacheGrid, and CompuStore

Ioan Raicu
Distributed Systems Laboratory
Computer Science Department
University of Chicago

DSL Seminar
November 1st, 2006

Analysis of Datasets: Data → Computation

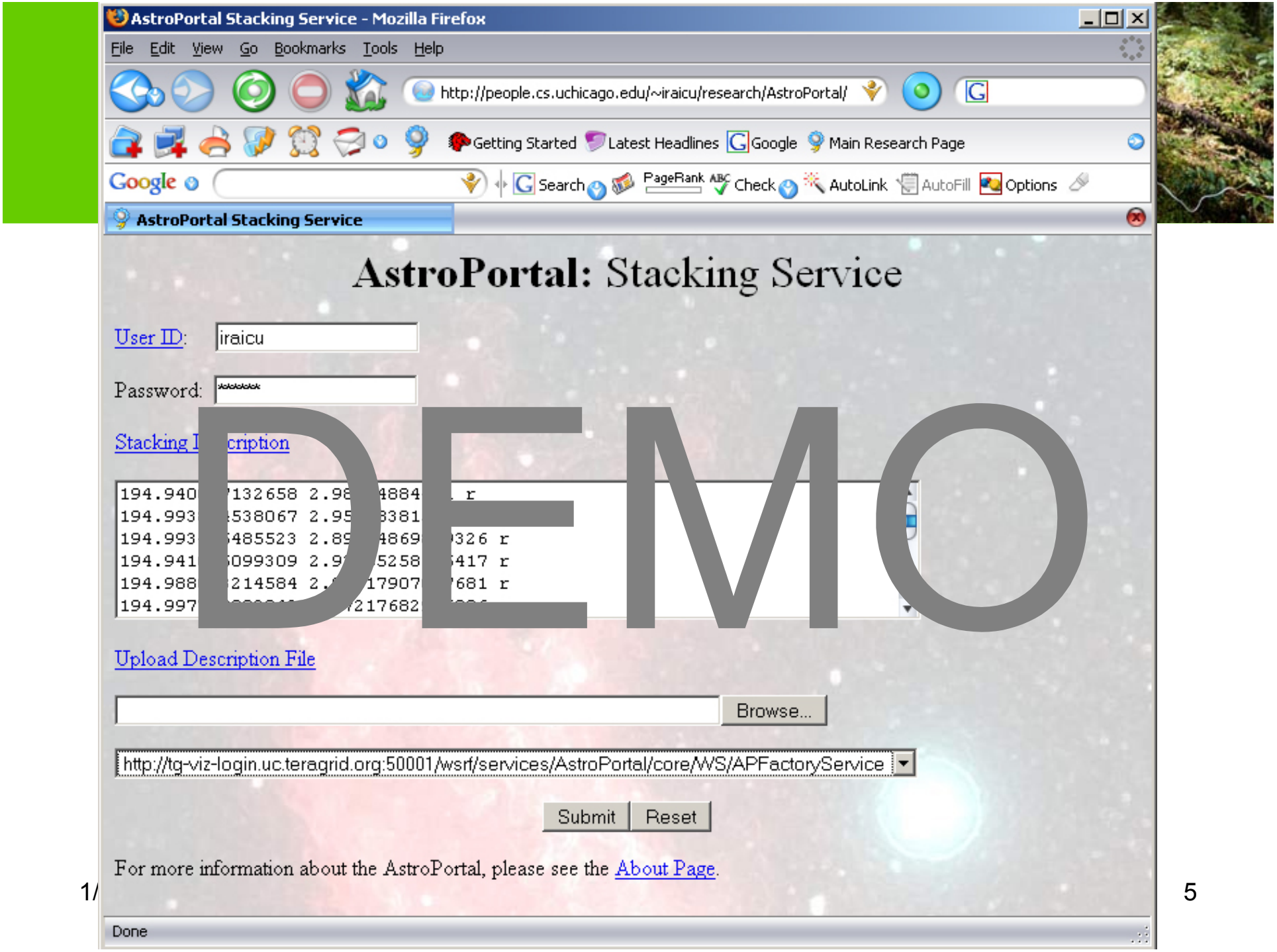


Dynamic & Distributed Analysis of Large Datasets

- Science Portals enable entire communities access to both compute and storage resources
  - Can enable the efficient analysis of large datasets
  - Move the computations to the data
- Potential Applications Characteristics
  - Large data sets
  - Large number of users
  - Relatively easy parallelization
- Applicable fields:
  - Astronomy
  - Medicine
  - Others

Astronomy Field

- Astronomy datasets (i.e. SDSS) are the crown-jewels
  - SDSS DR5
    - 1.5M images
      - 350M+ objects
      - 3TB compressed images (2MB x 1.5M)
      - 9TB raw images (6.1MB x 1.5M)
    - 100K worldwide potential users (100s of big users)
- Applications:
  - Stacking
  - Montage


AstroPortal Stacking Service - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://people.cs.uchicago.edu/~iraicu/research/AstroPortal/
Getting Started Latest Headlines Google Main Research Page
Search PageRank Check AutoLink AutoFill Options
AstroPortal Stacking Service
AstroPortal: Stacking Service
User ID: iraicu
Password:
Stacking I cription
194.940 7132658 2.98 4884 r
194.993 538067 2.95 3381
194.993 485523 2.89 4869 326 r
194.941 099309 2.9 5258 417 r
194.988 214584 2. 17907 681 r
194.997 217682
DEMO
Upload Description File
Browse...
http://tg-viz-login.uc.teragrid.org:50001/wsrf/services/AstroPortal/core/WS/APFactoryService
Submit Reset
For more information about the AstroPortal, please see the About Page.
Done

Stacking Results - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://people.cs.uchicago.edu/~iraicu/research/AstroPortal/results.htm

Getting Started Latest Headlines Google Main Research Page Des Plaines Public Library

Google Search PageRank Check AutoLink AutoFill Options

Stacking Results

AstroPortal: Stacking Service Results

User ID: iraicu
Password: ******
Stacking Description: stacking_description.txt
Stacking Size: 20
AstroPortal Web Service Location: http://tg-viz-login.uc.teragrid.org:50001/wsrf/services/AstroPortal/core/WS/APFactoryService

RESULT:

Size: 43 KB
Dimensions: 100x100 pixels
Download result: stacked_result.fit

Time to complete Stacking: 5.164 seconds
Number of physical resources utilized: 16
Number of Stackings completed successful: 18
Number of Star Objects not found in the SDSS dataset: 1
List of Star Objects [ra, dec, band] not found:

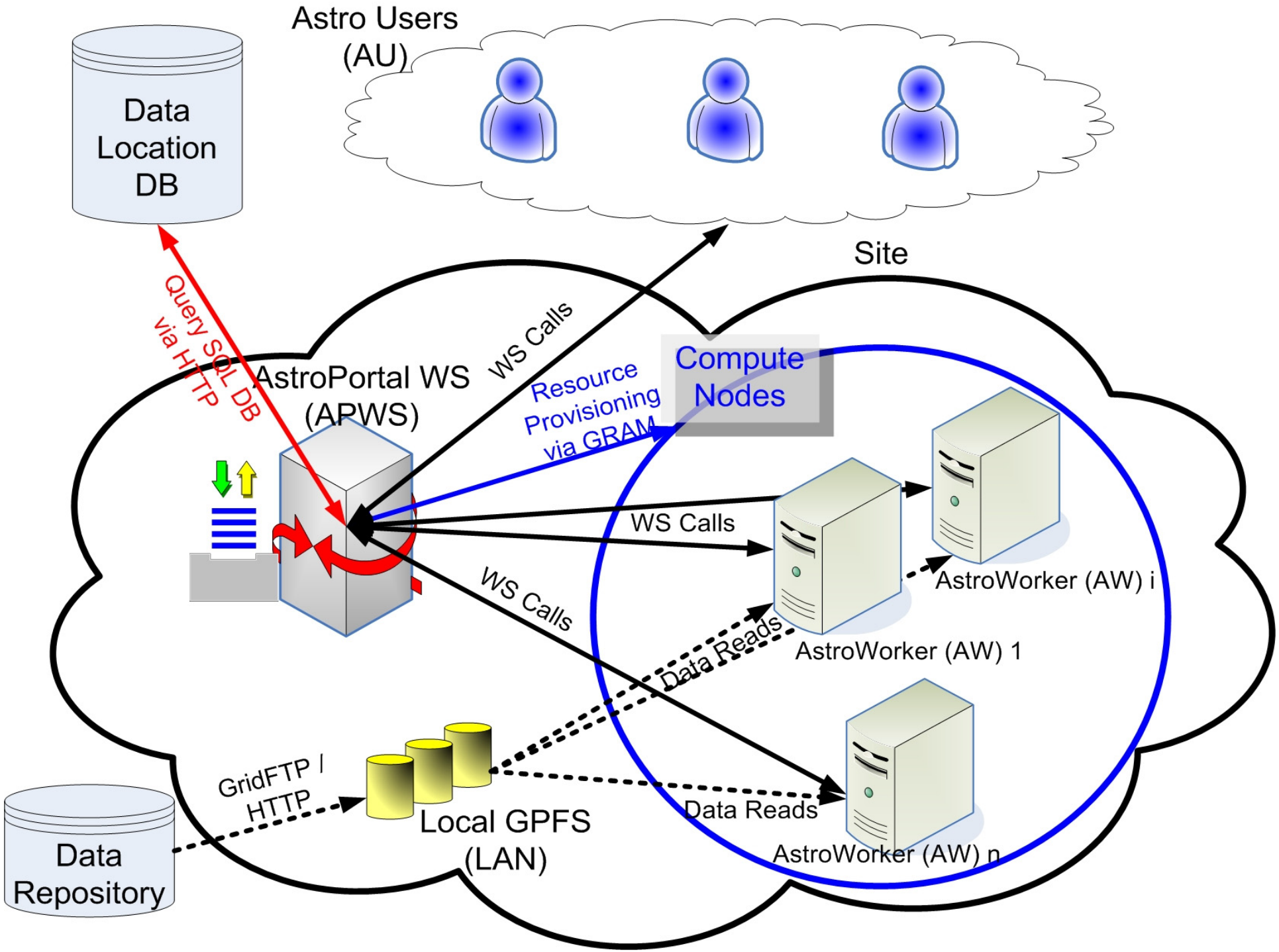
- [194.969060213455, -13.90189344168167, r]

Number of Data Objects not found in the data cache: 1
List of Data Objects {[ra, dec, band] filename [x_coord x y_coord]} not found:

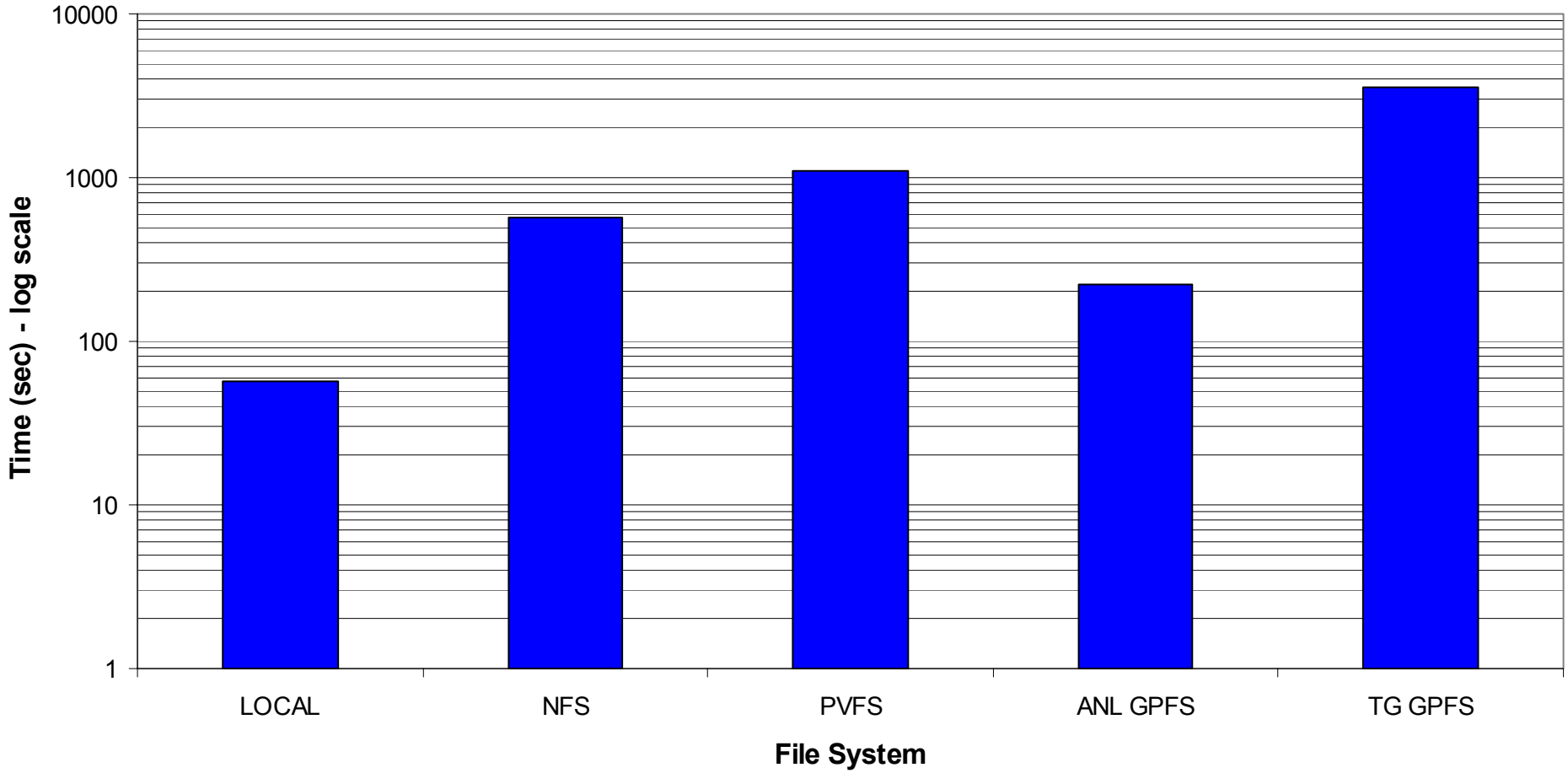
- {[194.969705877549, 2.93855950426612, r] /disks/scratchgpfs1/iraicu/sdss.gz/das.sdss.org/DR4/data/imaging/752/40/corr/6/fpC-000752-r6-0245.fit.gz [0 x 0]}

To start a new stacking, go back to the main Stacking Service.

Done


Astro Users
(AU)
Data
Location
DB
Site
Query SQL DB
via HTTP
AstroPortal WS
(APWS)
WS Calls
Resource
Provisioning
via GRAM
Compute
Nodes
WS Calls
WS Calls
AstroWorker (AW) i
AstroWorker (AW) 1
Data Reads
GridFTP |
HTTP
Local GPFS
(LAN)
Data Reads
AstroWorker (AW) n
Data
Repository

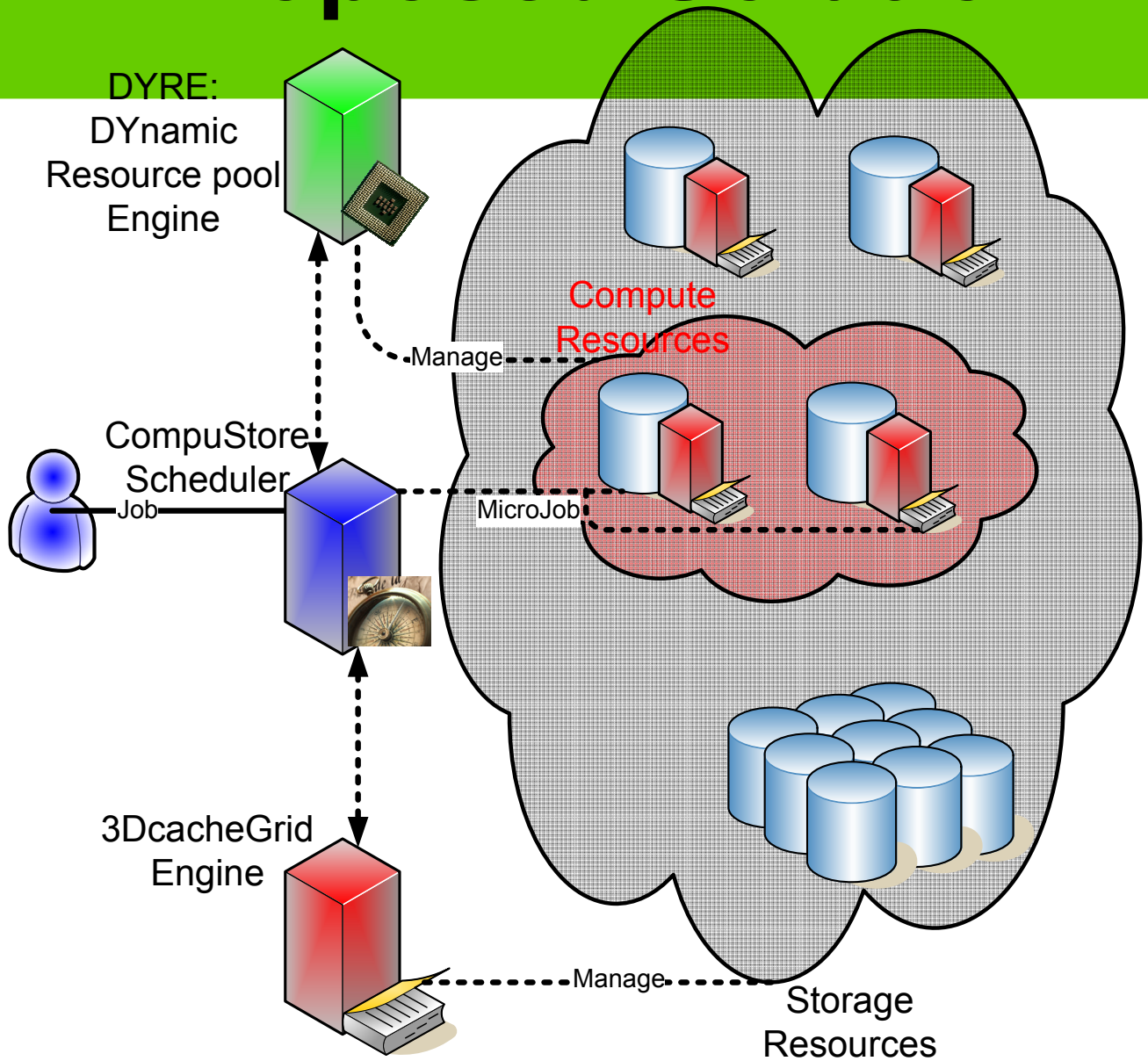
O(100K)


10000
1000
100
10
1
Time (sec) - log scale
LOCAL
NFS
PVFS
ANL GPFS
TG GPFS
File System

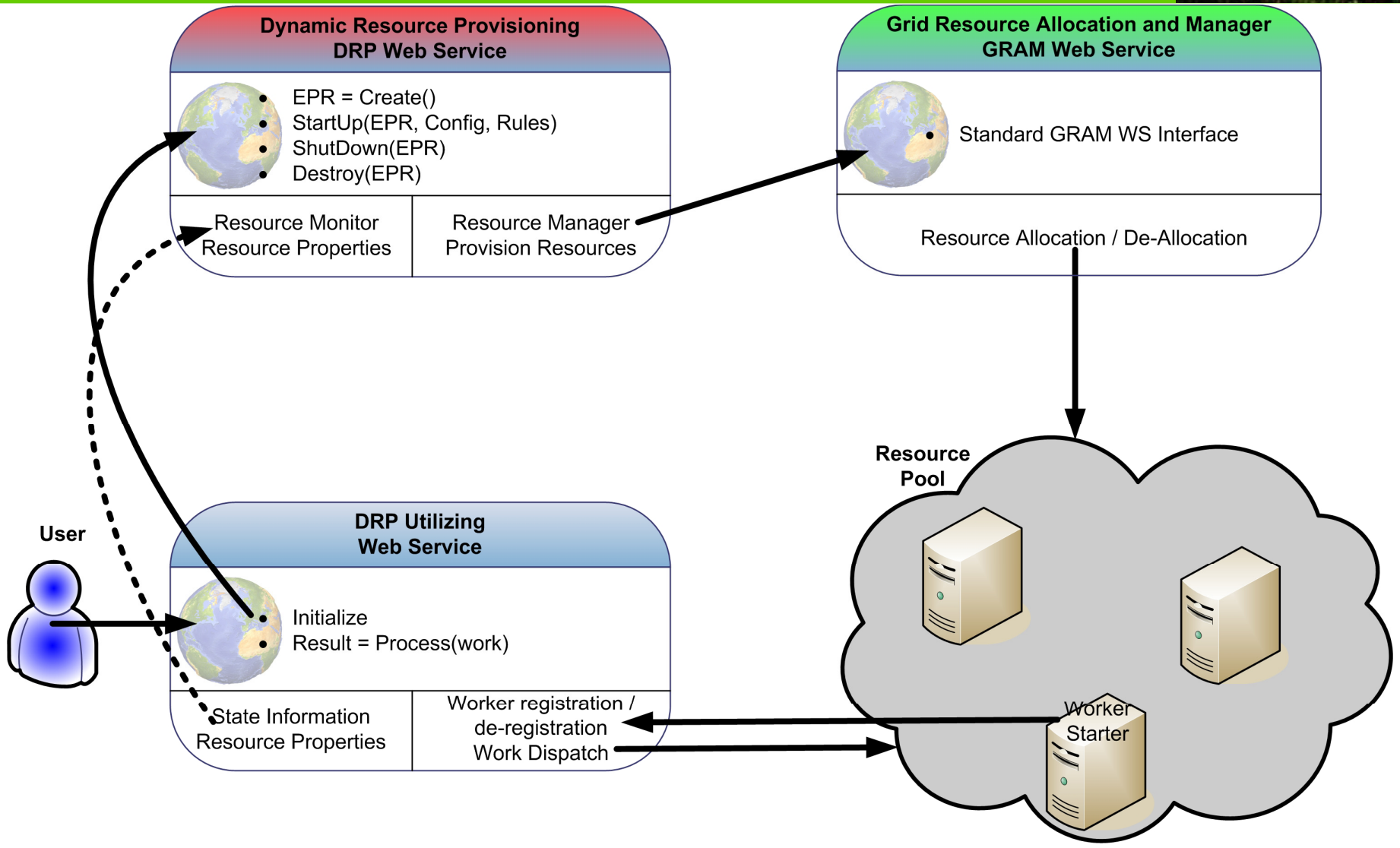
Open Research Questions

- Data Resource management
  - Data set distribution among various storage resources
  - Data placement based on past workloads and access patterns
    - Caching strategies: LRU, FIFO, popularity, …
    - Replication strategies to meet a desired QoS
  - Data management architectures
- Compute Resource management
  - Resource Provisioning
  - Harness entire TeraGrid pool of resources
  - Workload management, moving the work vs. moving the data
  - Distributed resource management between various sites
  - Scheduling of computations close to data

Proposed Solution



DRP: Dynamic Resource Provisioning



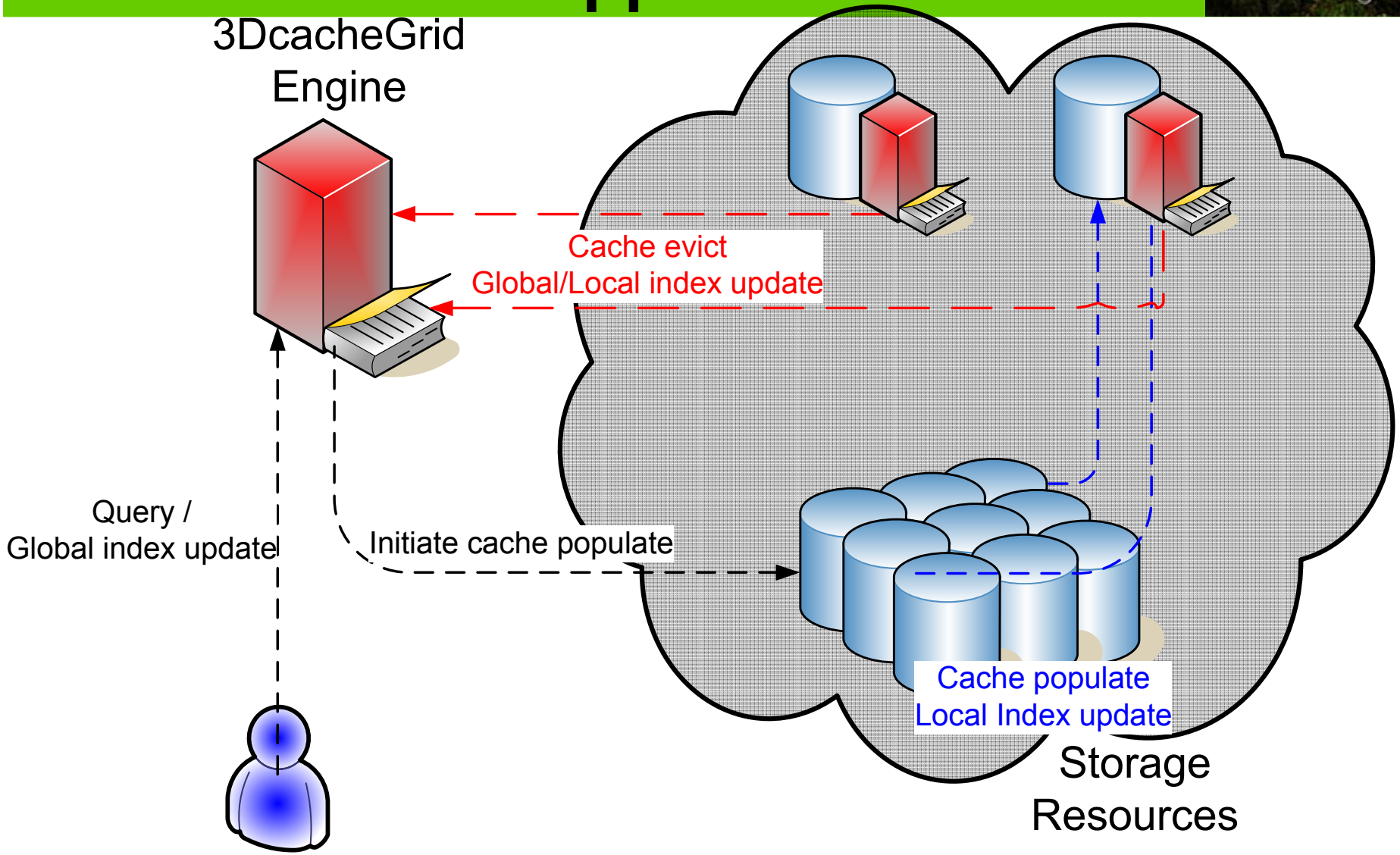
DRP Advantages

- Allows for finer grained resource management, including the control of priorities and usage policies
- Optimize for the grid user's perspective: reduces delays on per job scheduling by utilizing pre-reserved resources
- Increased resource utilization (on the surface)
- Opens the possibility to customize the resource scheduler per application basis
  - use of both data resource management and compute resource management information for more efficient scheduling
- Reduced complexity to the application developer

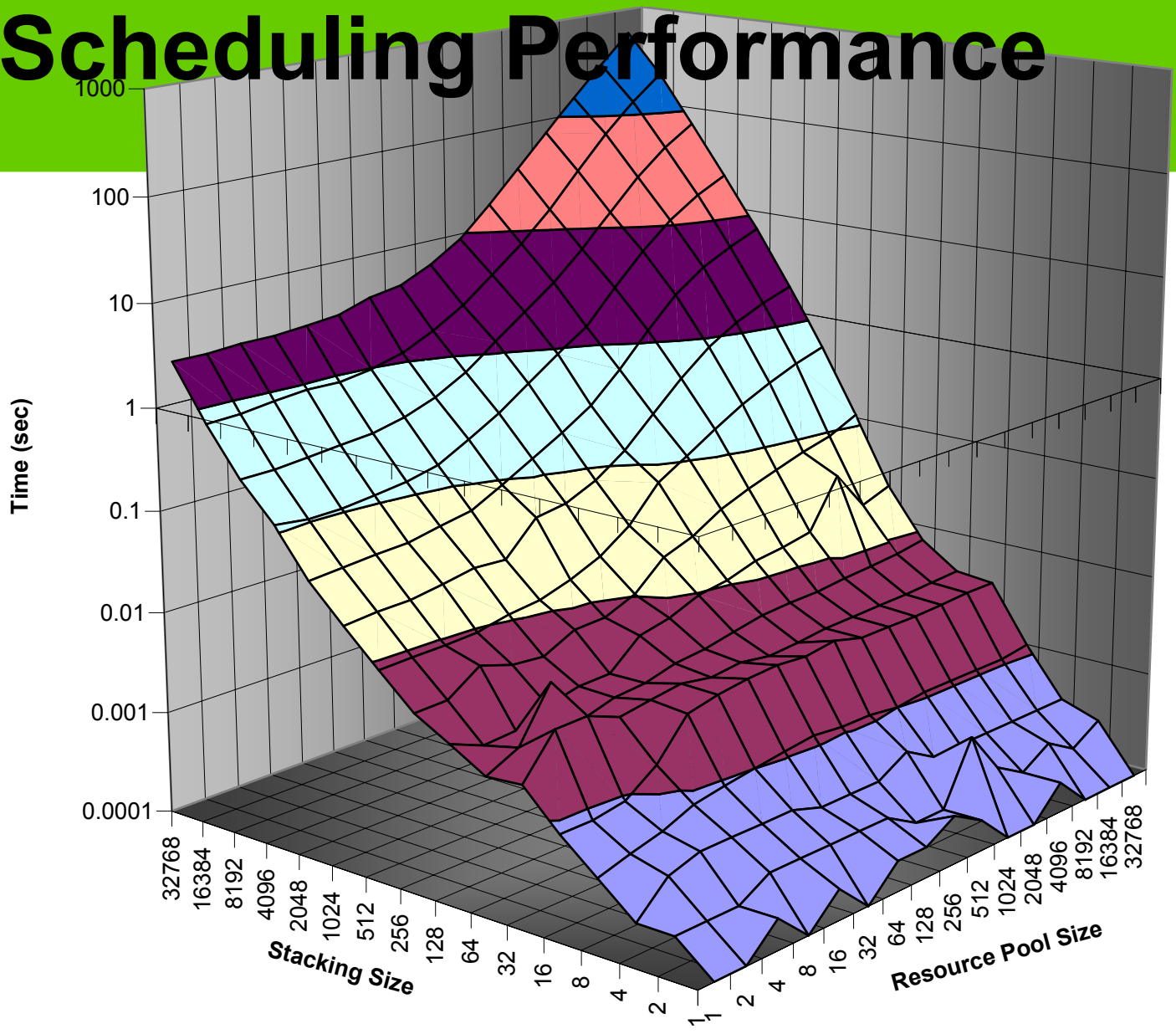
DRP Disadvantages

- All jobs submitted by different members need to map to the same user
- Initial startup overhead
- Work could be halted unfinished when the original time lease on a particular resource expires if the time lease not being exposed to the work dispatcher
- Underutilization of raw resources

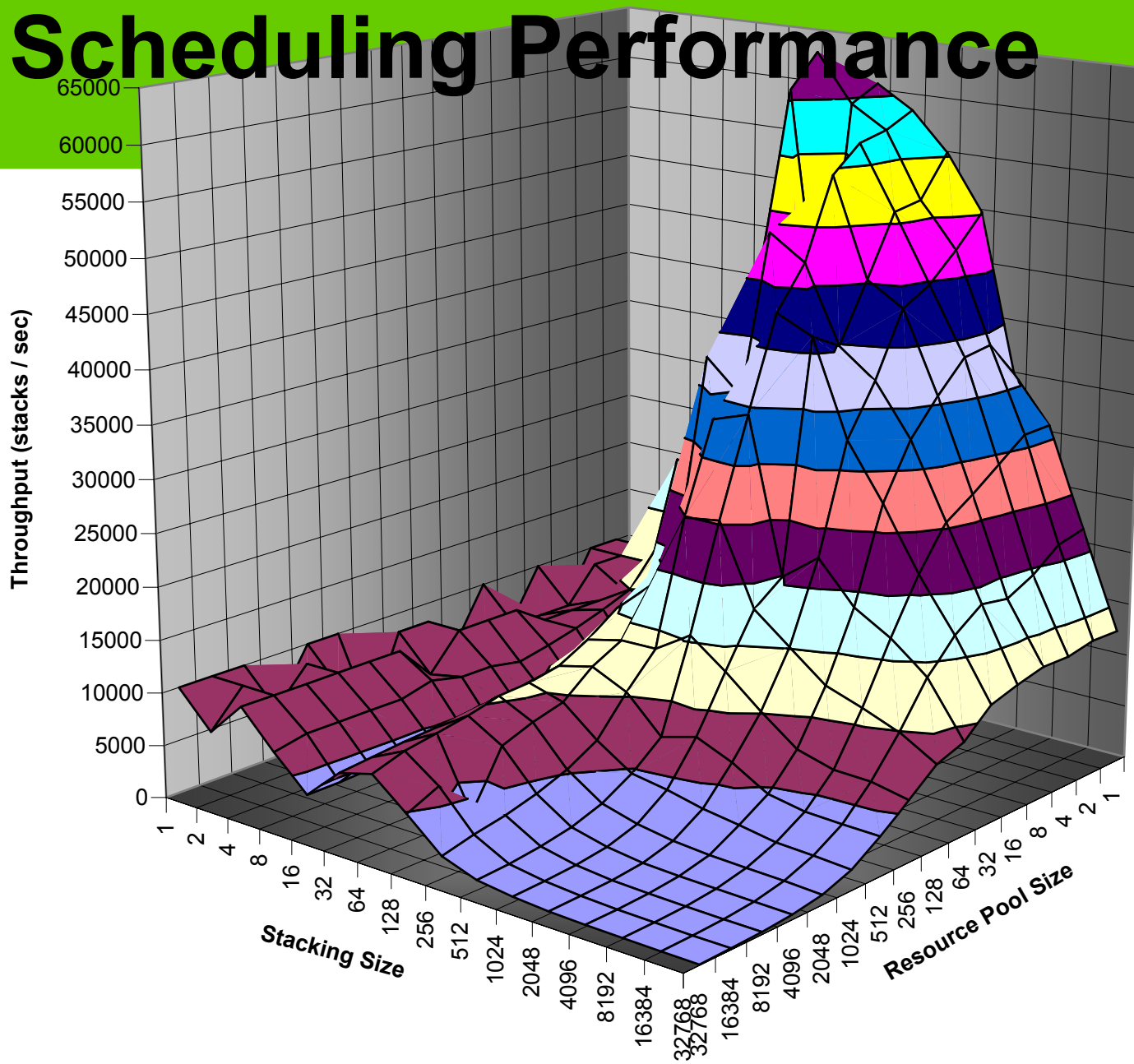
3DcacheGrid Engine: Dynamic Distributed Data cache for Grid Applications



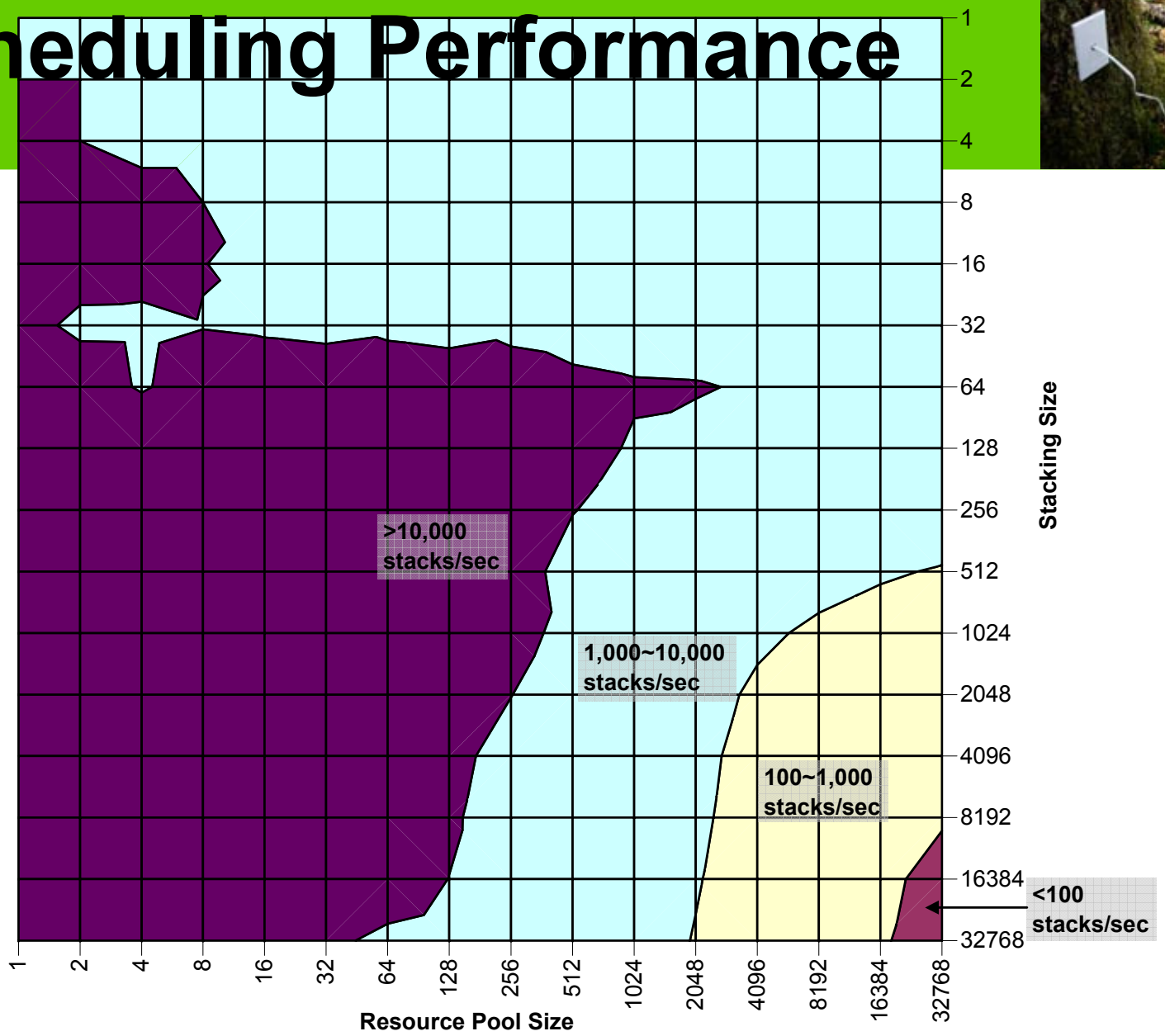
Data Management & Scheduling Performance



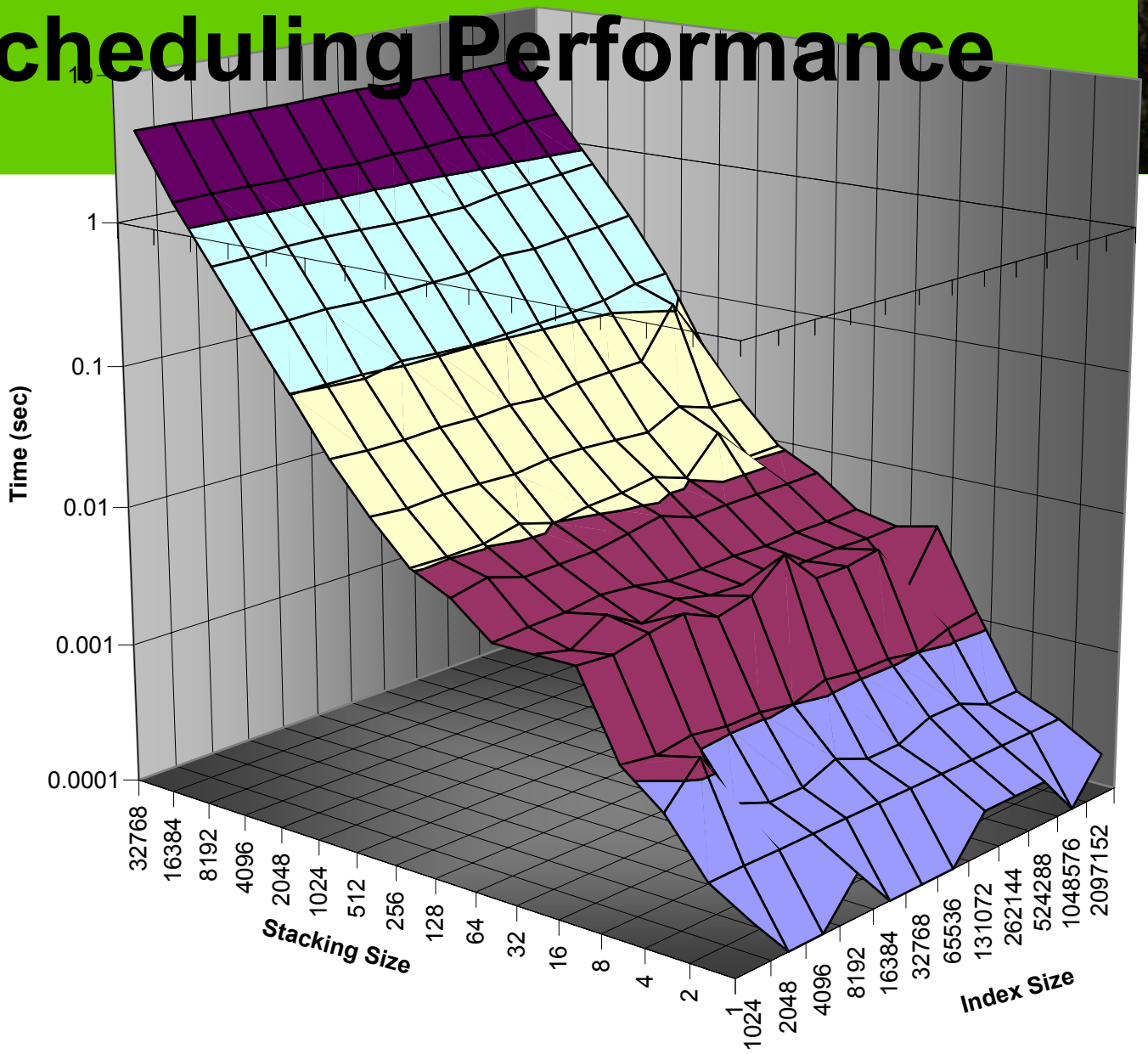
Data Management & Scheduling Performance



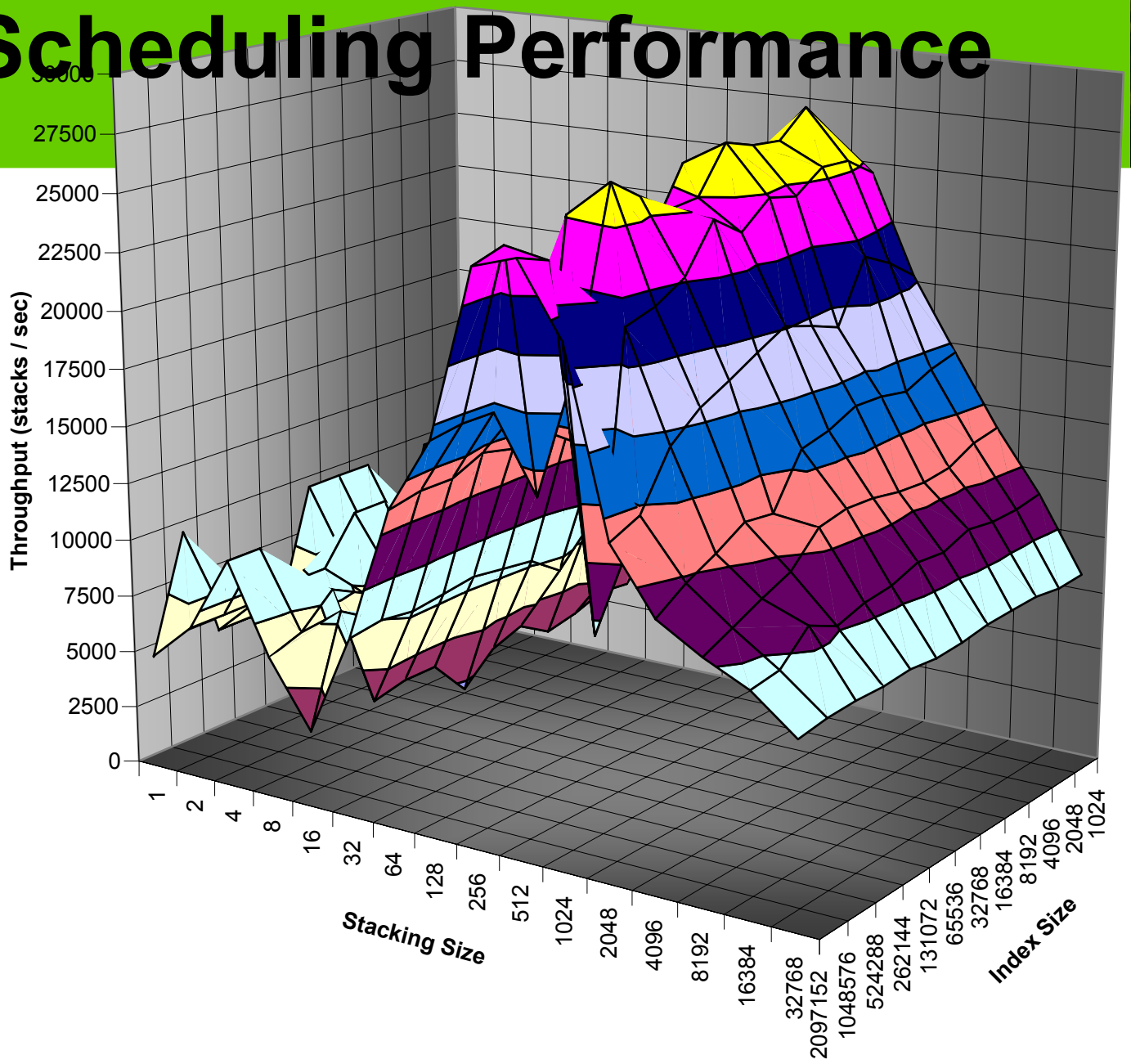
Data Management & Scheduling Performance



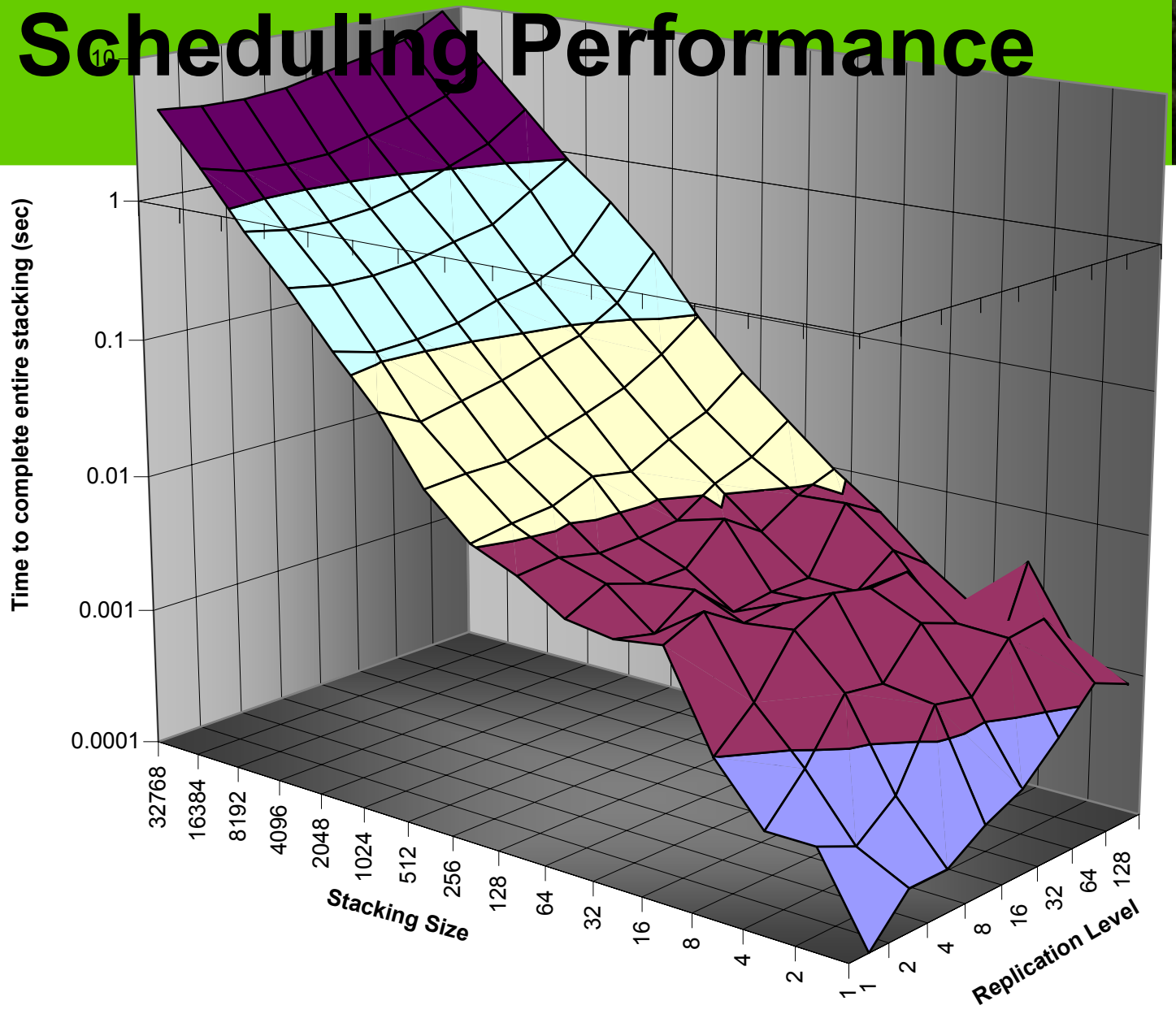
Data Management & Scheduling Performance



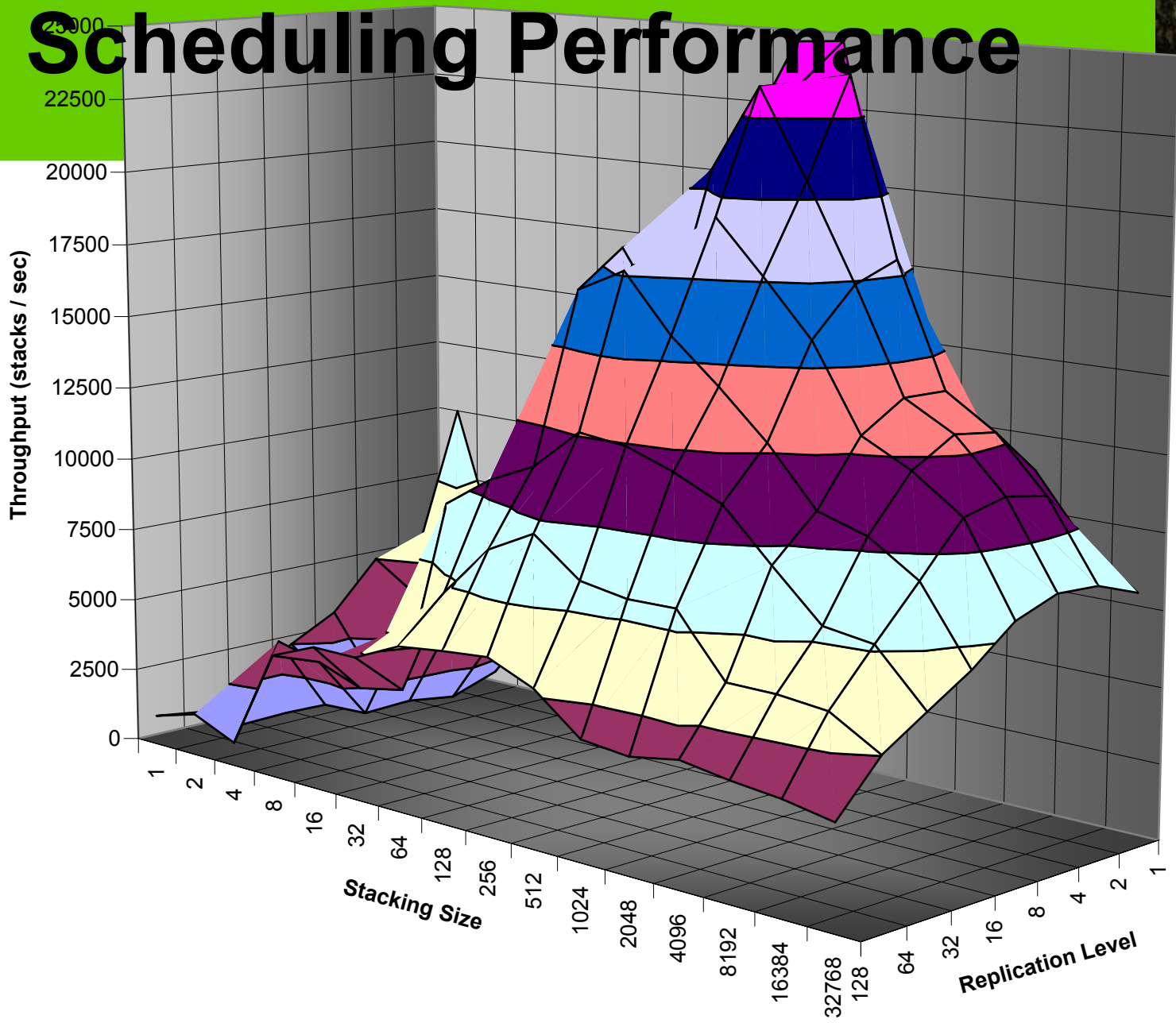
Data Management & Scheduling Performance



Data Management & Scheduling Performance



Data Management & Scheduling Performance



Data Management & Scheduling Performance Conclusions

- Stacking size: less than 32K (although another order of magnitude probably won't pose any performance risks)
- Resource pool size: less than 1000 resources might offer decent performance if there is the replication level remains low, but for higher orders of replication, less than 100 resources are recommended
- Index Size: 2M~10M depending on the level of replication using a 1.5GB Java heap; larger index sizes could be supported linearly without sacrificing performance by increasing the Java heap size (needing more physical memory and possibly a 64 bit JVM environment)
- Replication Level: less than 128 replicas (although more could be supported as long as the dataset size remains relatively fixed)
- Resource Capacity: 100GB of local storage per resource (this could be increased, but its unclear what the performance effects would be)

Questions?

- More information: http://people.cs.uchicago.edu/~iraicu/research/
- Related materials and further readings:
  - Ioan Raicu, Ian Foster, Alex Szalay, Gabriela Turcu. "***AstroPortal: A Science Gateway for Large-scale Astronomy Data Analysis***", TeraGrid Conference 2006, June 2006.
  - Alex Szalay, Julian Bunn, Jim Gray, Ian Foster, Ioan Raicu. "***The Importance of Data Locality in Distributed Computing Applications***", NSF Workflow Workshop 2006.
  - Ioan Raicu, Ian Foster, Alex Szalay. "***Harnessing Grid Resources to Enable the Dynamic Analysis of Large Astronomy Datasets***", SuperComputing 2006.
  - Ioan Raicu. "Harnessing Grid Resources to Enable the Dynamic Analysis of Large Astronomy Datasets", NASA Ames Research Center GSRP Proposal, funded 10/2006 – 9/2007