THE UNIVERSITY OF CHICAGO

AstroPortal: A Science Gateway for Large-scale Astronomy Data Analysis

Ioan Raicu
Distributed Systems Laboratory
Computer Science Department
University of Chicago

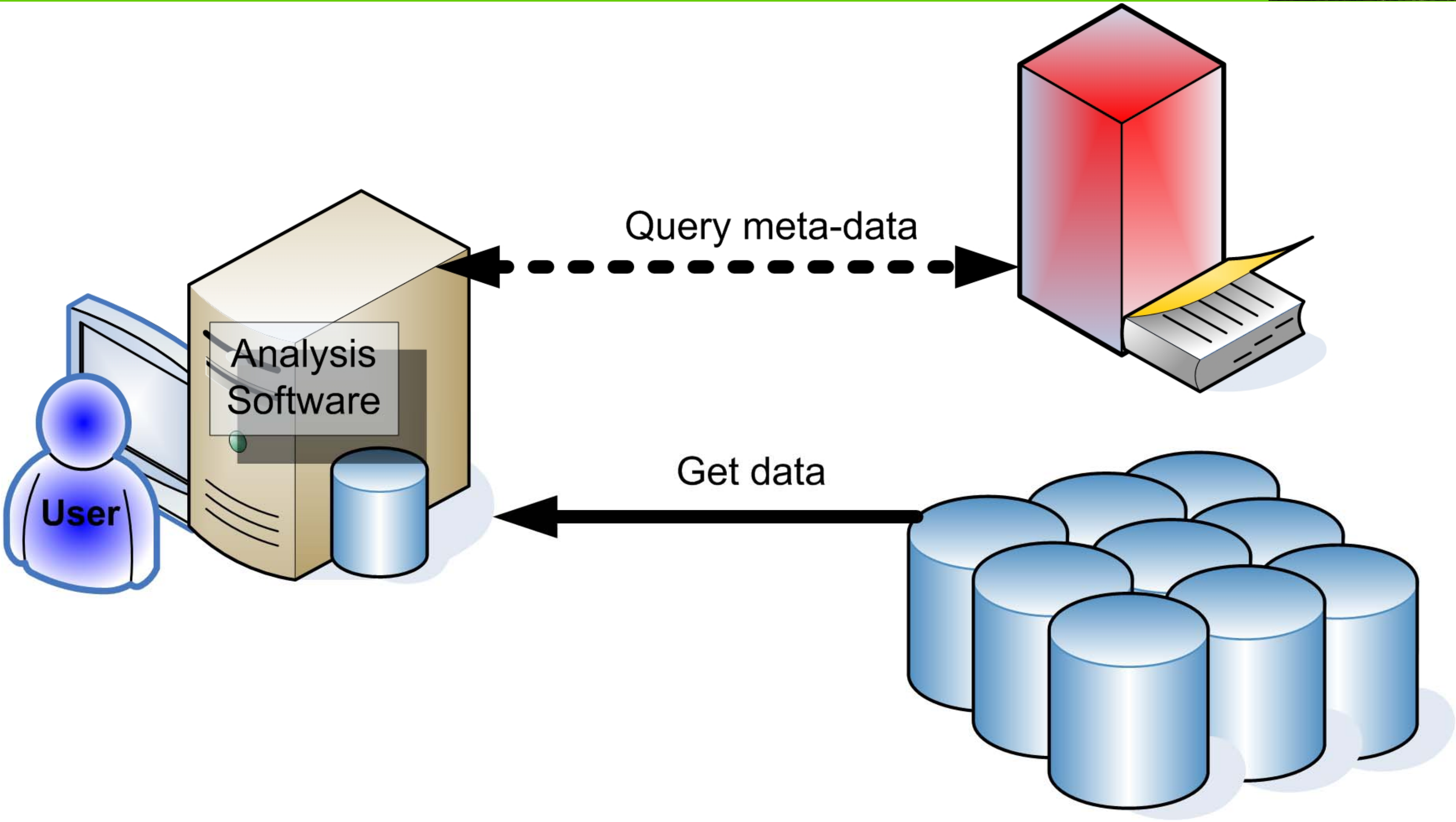
Joint work with:
Ian Foster: Univ. of Chicago, CS & Argonne National Laboratory, MCS
Alex Szalay: Johns Hopkins University, Dept. of Physics and Astronomy
Gabriela Turcu: Univ. of Chicago, CS

Funded by:
NSF TeraGrid: June 2005 – September 2006
NASA Ames Research Center GSRP: October 2006 – September 2007

AstroGrid 2007 Meeting
February 12th, 2007

Argonne NATIONAL LABORATORY

Analysis of Datasets: Data → Computation



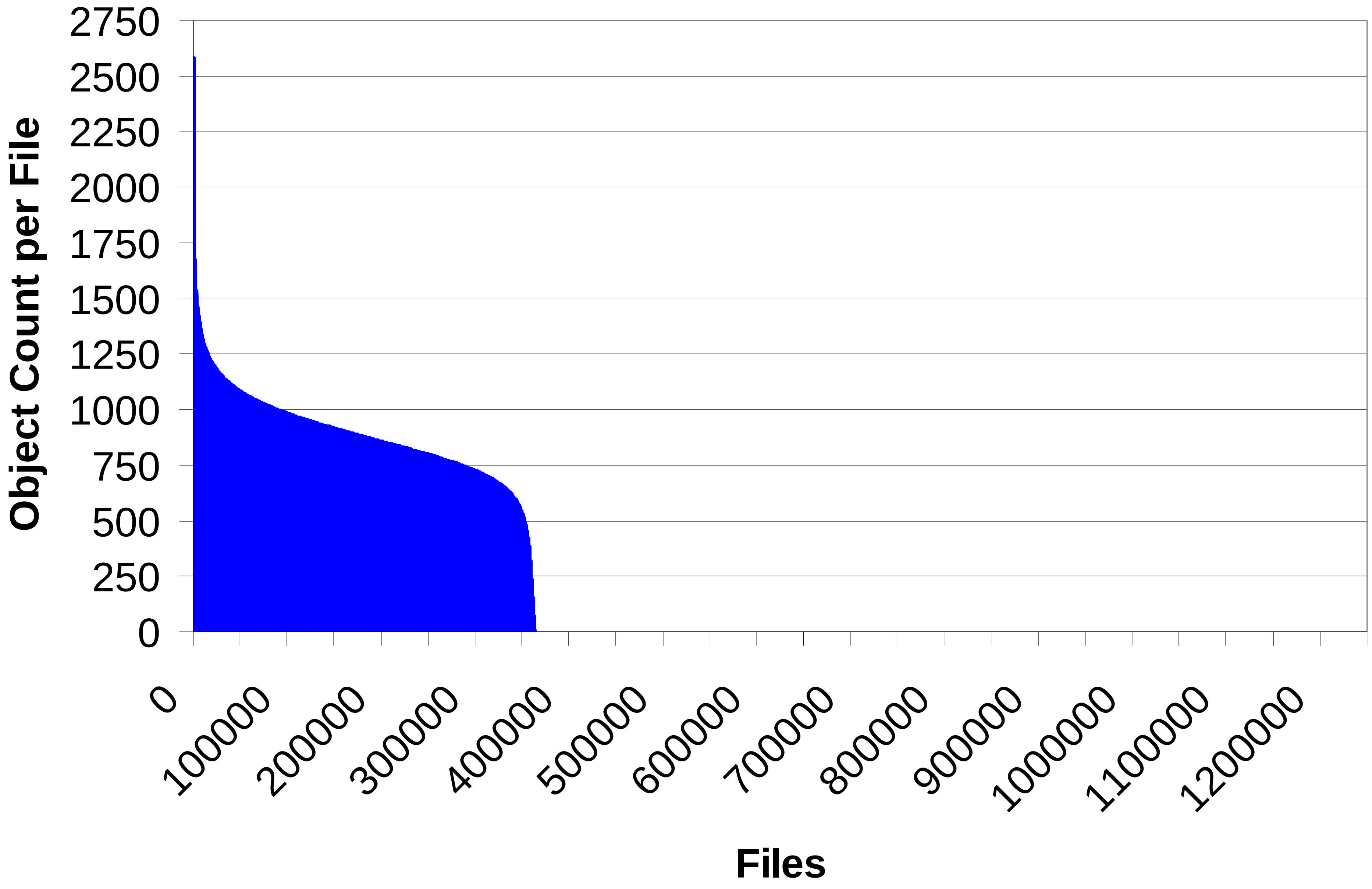
Dynamic & Distributed Analysis of Large Datasets

- 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

Object Distribution SDSS DR4



AstroPortal Stacking Service - Mozilla Firefox

File Edit View History Bookmarks Tools Help iraicu

http://s8.uchicago.edu:8080/AstroPortal/index.jsp

AstroPortal: Stacking Service

What is the AstroPortal?

UserID: iraicu **Password:** ********

Need an account or forgot your password, click here.

Stacking Description (click here for details)
Upload file: Browse...

Copy and Paste:

```
194.940047132658 2.98364884441 i
194.993834538067 2.95438381572631 u
194.993436485523 2.89844869849326 z
194.941075099309 2.93405258125417 g
194.988003214584 2.910179 [illegible]
194.940047132658 2.983648 [illegible]
194.993834538067 2.954383 [illegible] 2631 u
194.993436485523 2.898448 [illegible] 9326 z
194.941075099309 2.934052 [illegible] 5417 g
194.988003214584 2.910179 [illegible] 7681 r
194.940047132658 2.983648 [illegible] 1 i
```

Height 100 **Width** 100

AstroPortal Web Service Loca[illegible]
http://tg-viz-login2.uc.teragrid.org:[illegible]/AstroPortal/core/WS/[illegible]

Submit Reset

Understanding the results, click here for details.

Understanding any errors that might occur, click here for details.

Please report any problems, issues, or comments to Ioan Raicu.

Frequently Asked Questions (FAQ)

Done

DEMO

AstroPortal Stacking Service - Mozilla Firefox

File Edit View History Bookmarks Tools Help iraicu

http://s8.uchicago.edu:8080/AstroPortal/results.jsp

Time (sec)	Queued Stackings	Active Stackings	Completed Stackings	Submitted Stackings	Completed(%)	Global Queued Stackings	Global Active Stackings	Global Active Resources
6.808	1328	672	0	2000	0%	1328	672	32
9.068	1328	567	147	2000	7%	1328	546	32
10.151	1328	420	315	2000	15%	1328	357	32
10.735	1265	189	630	2000	31%	1265	105	32
11.898	1097	315	672	2000	33%	1097	231	32
12.048	1076	336	672	2000	33%	1076	252	32
13.27	845	567	672	2000	33%	845	483	32
13.431	803	609	672	2000	33%	803	525	32
14.68	698	714	672	2000	33%	698	630	32
14.832	677	735	672	2000	33%	677	651	32
16.143	656	609	819	2000	40%	656	525	32
[illegible]	656	378	1071	[illegible]	[illegible]	656	273	32
[illegible]	509	[illegible]	1281	[illegible]	64%	509	231	32
[illegible]	257	[illegible]	1344	[illegible]	67%	[illegible]	399	32
[illegible]	68	[illegible]	1344	2000	67%	[illegible]	588	32
[illegible]	0	[illegible]	1449	2000	72%	[illegible]	551	32
[illegible]	[illegible]	[illegible]	1685	2000	84%	[illegible]	315	32
[illegible]	[illegible]	189	1916	[illegible]	95%	[illegible]	84	32
[illegible]	[illegible]	147	1958	[illegible]	97%	0	42	32
[illegible]	[illegible]	105	2000	[illegible]	[illegible]	0	0	32

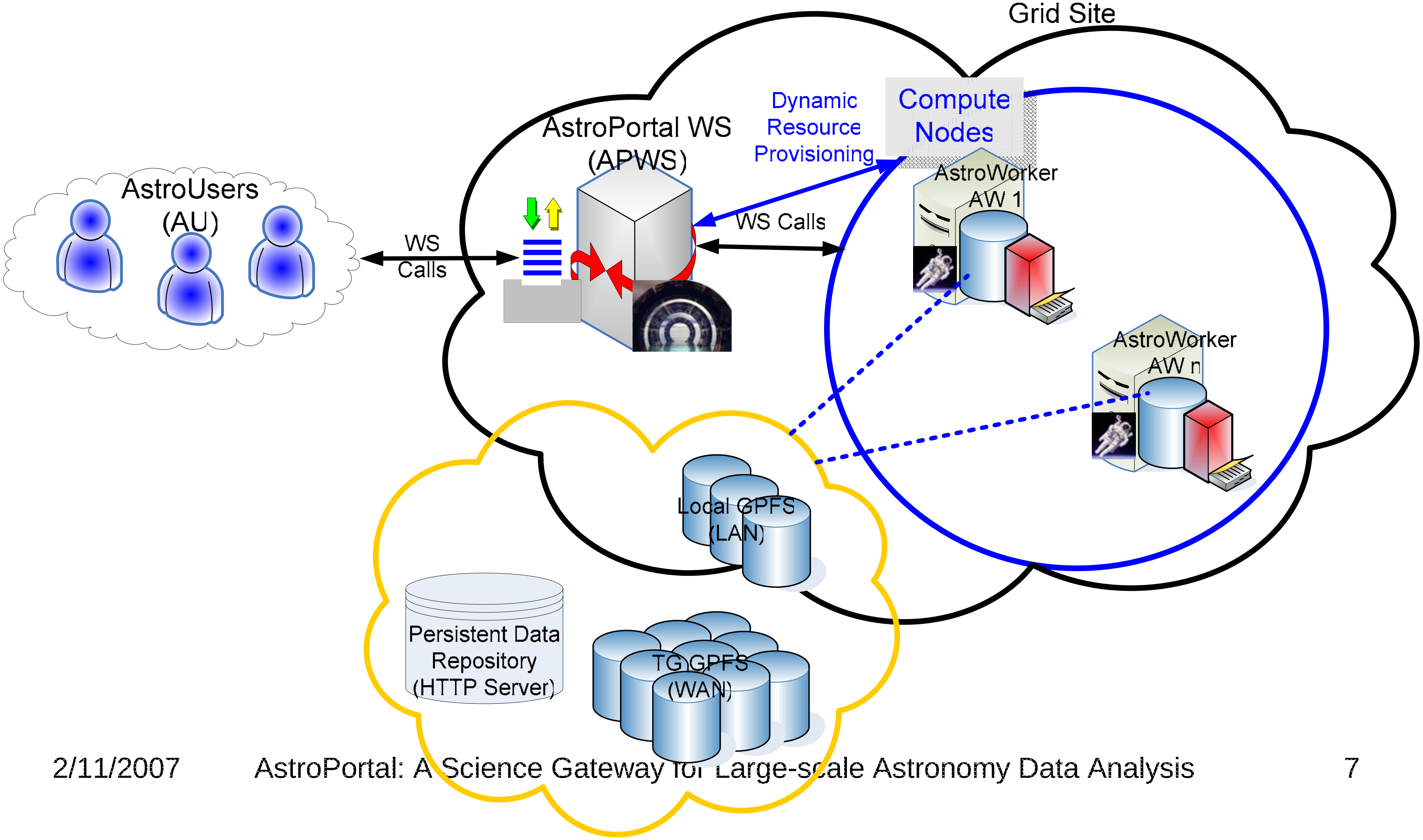
Received results in 29508.0ms
Total number of images requested 2000
2000 actual images stacked
0 requested stackings not performed

Download fit image

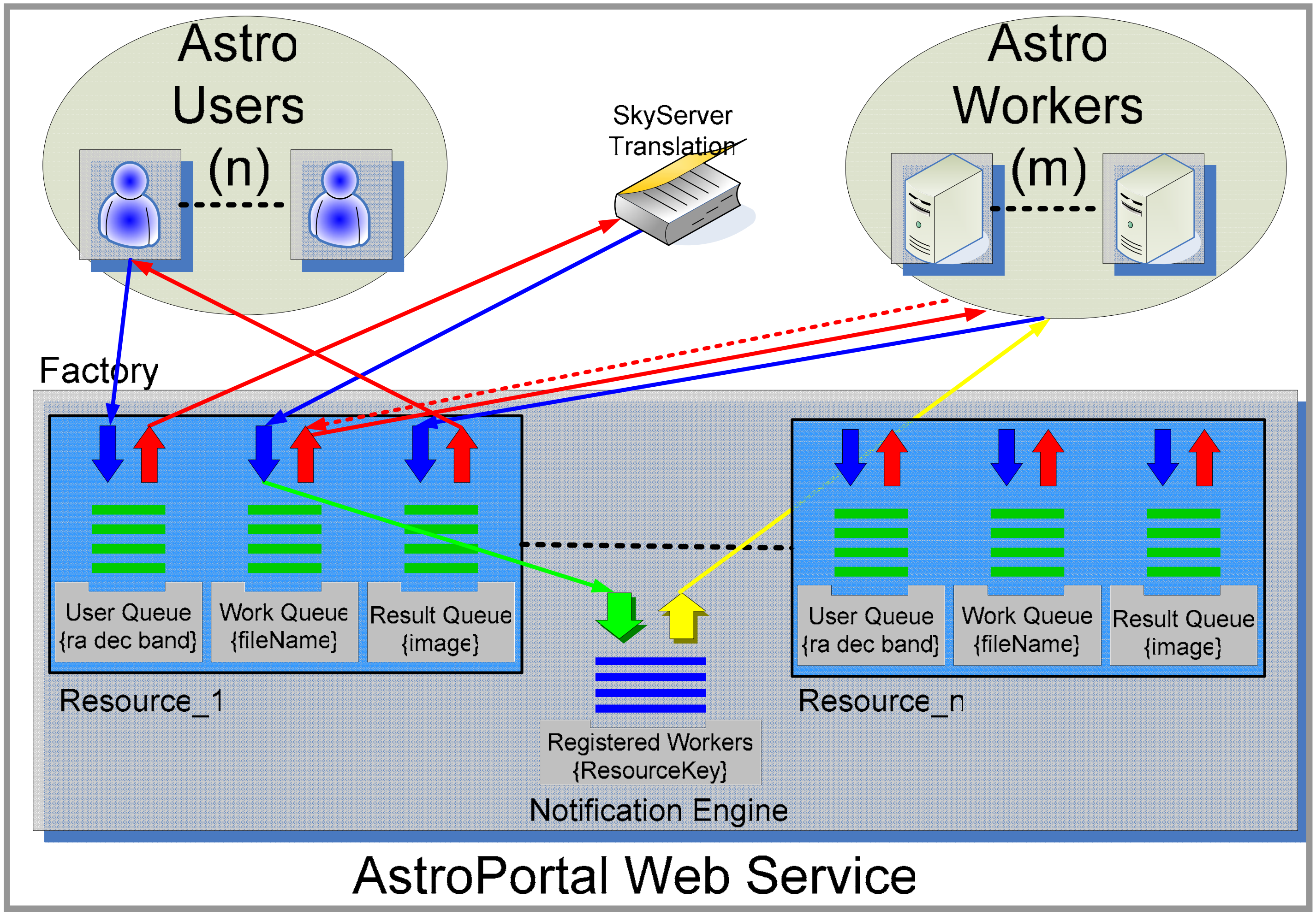
New search

Done

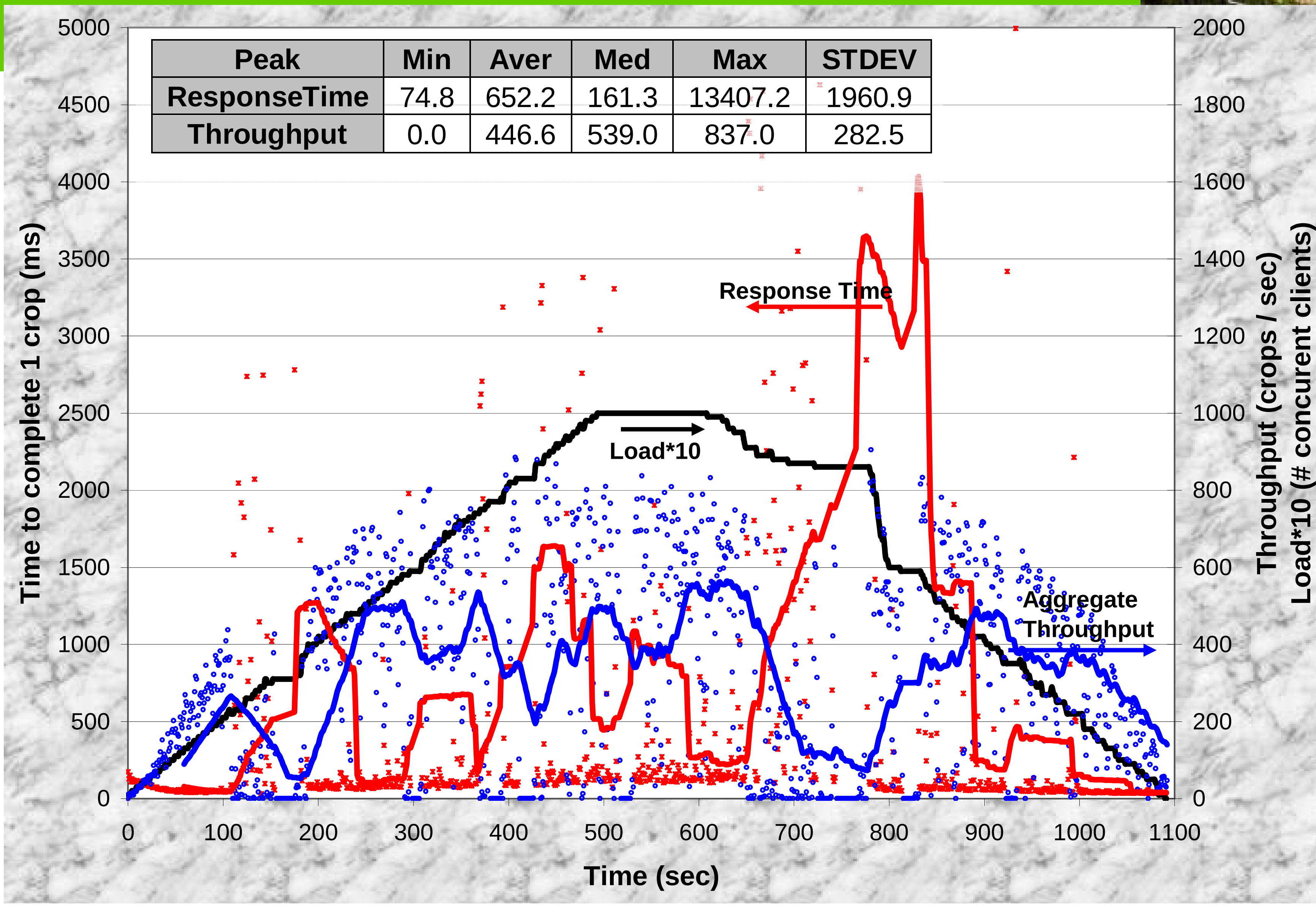
Architecture Overview



AstroPortal Web Service

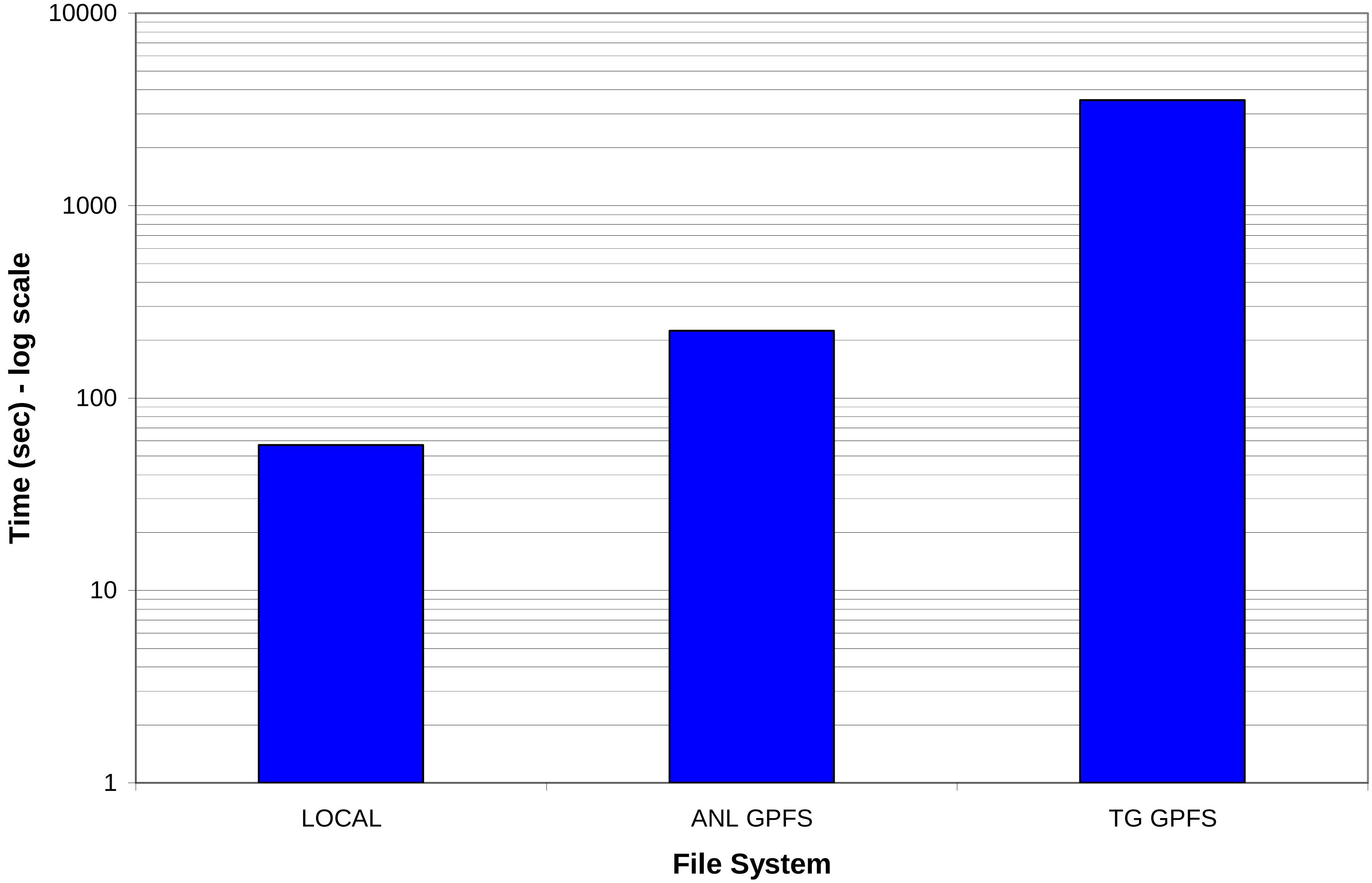


Raw Cutout Performance LAN GPFS in GZ Format

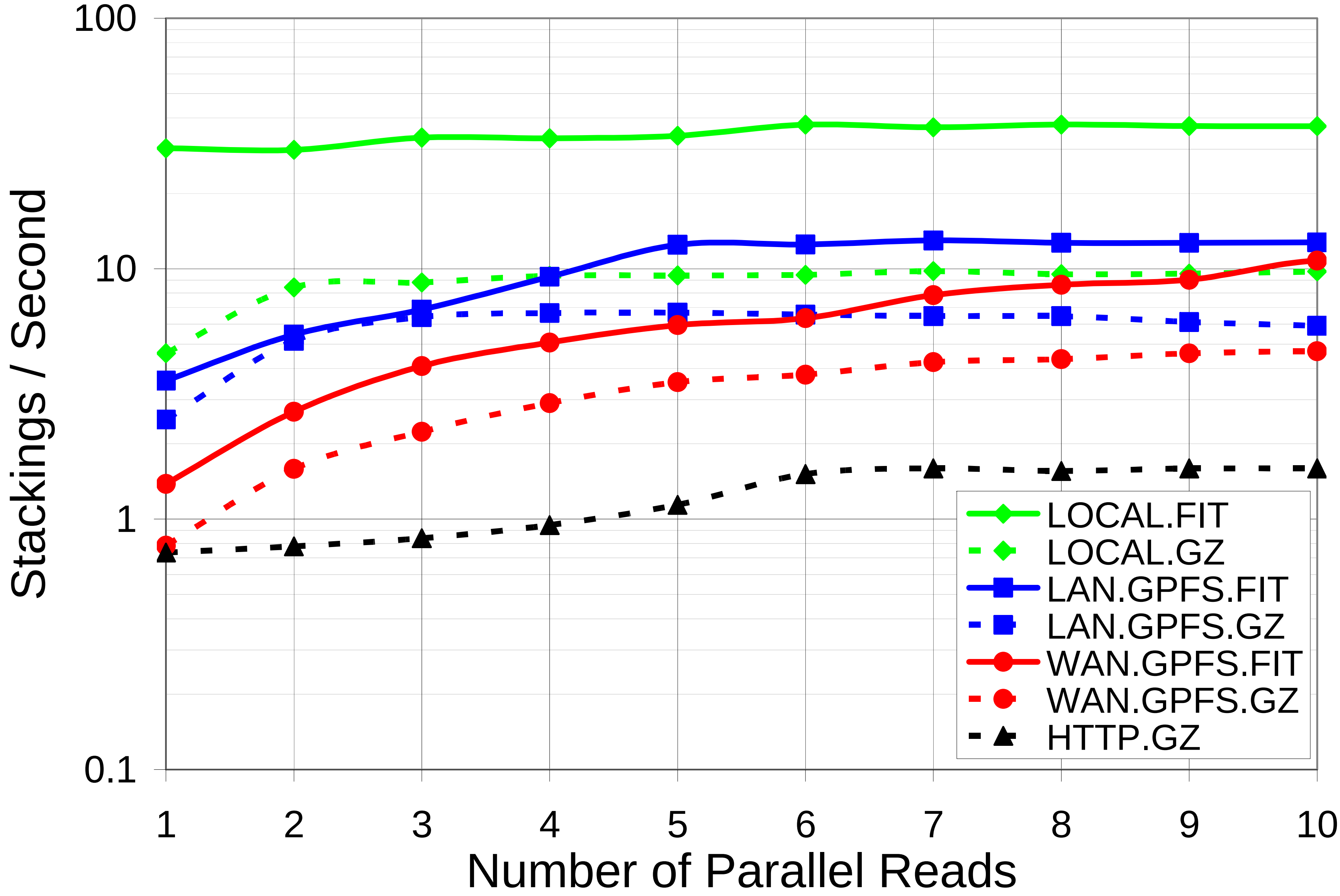


Peak	Min	Aver	Med	Max	STDEV
ResponseTime	74.8	652.2	161.3	13407.2	1960.9
Throughput	0.0	446.6	539.0	837.0	282.5

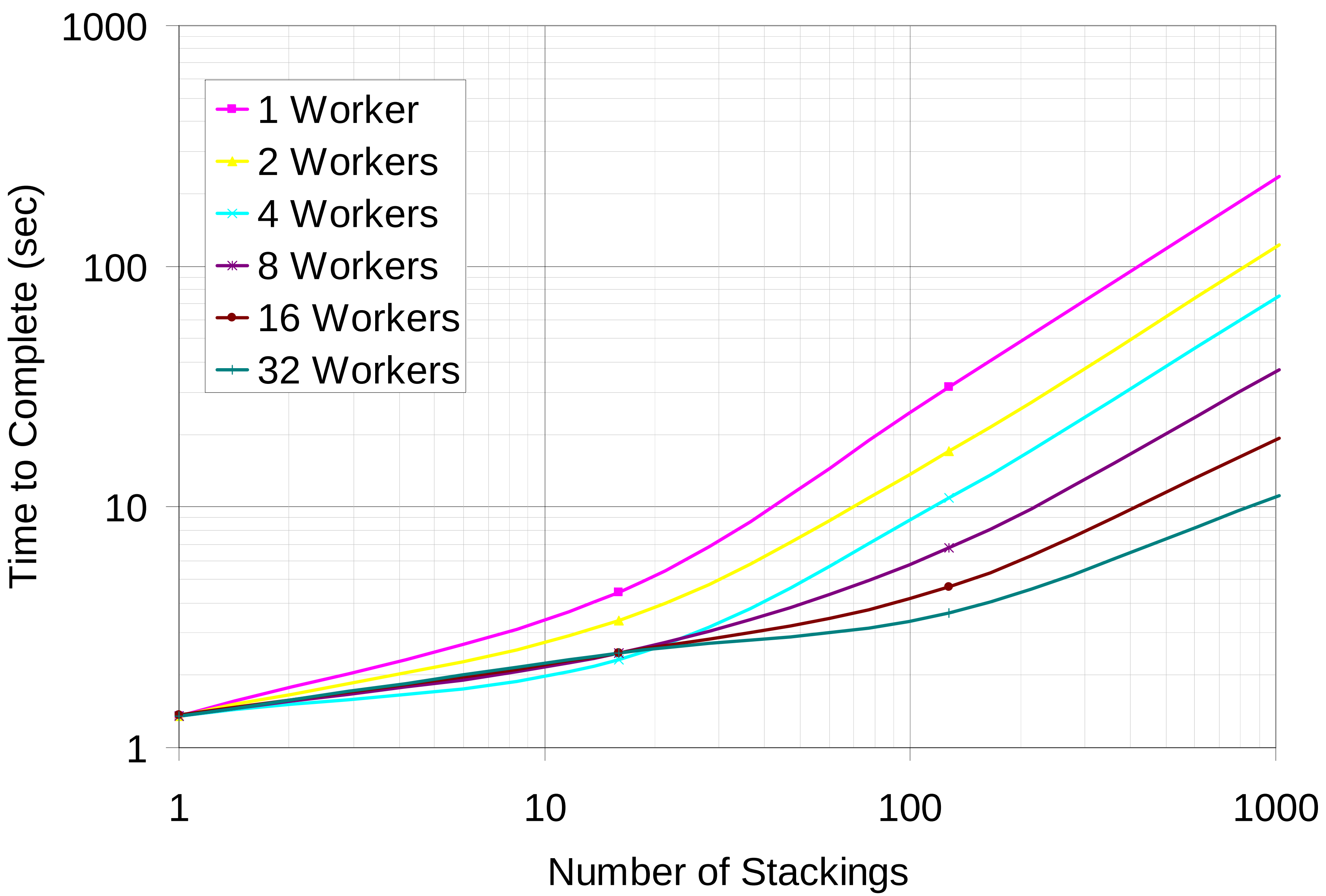
O(100K) Cutouts



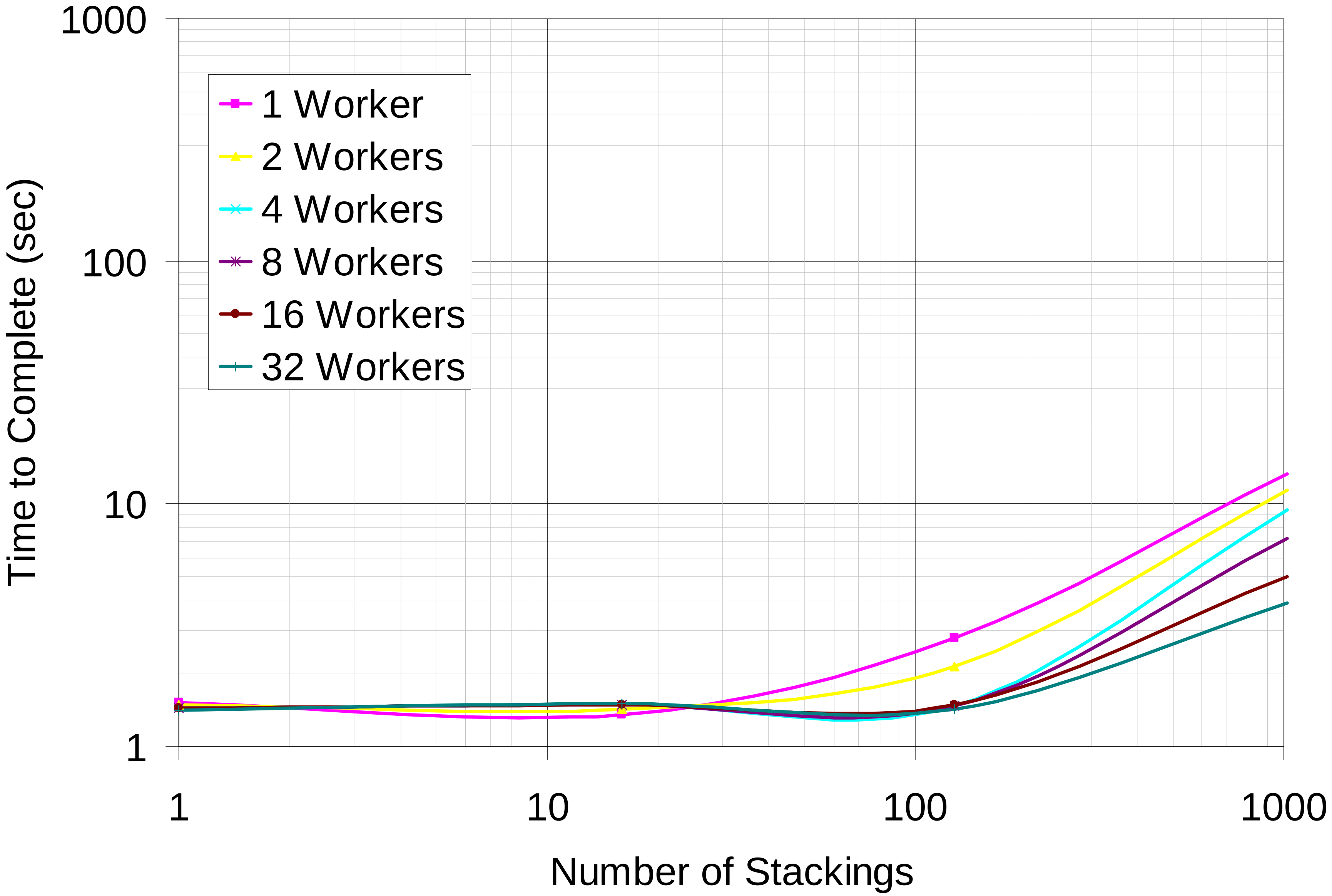
Raw Stacking
1 Worker – Multiple Threads



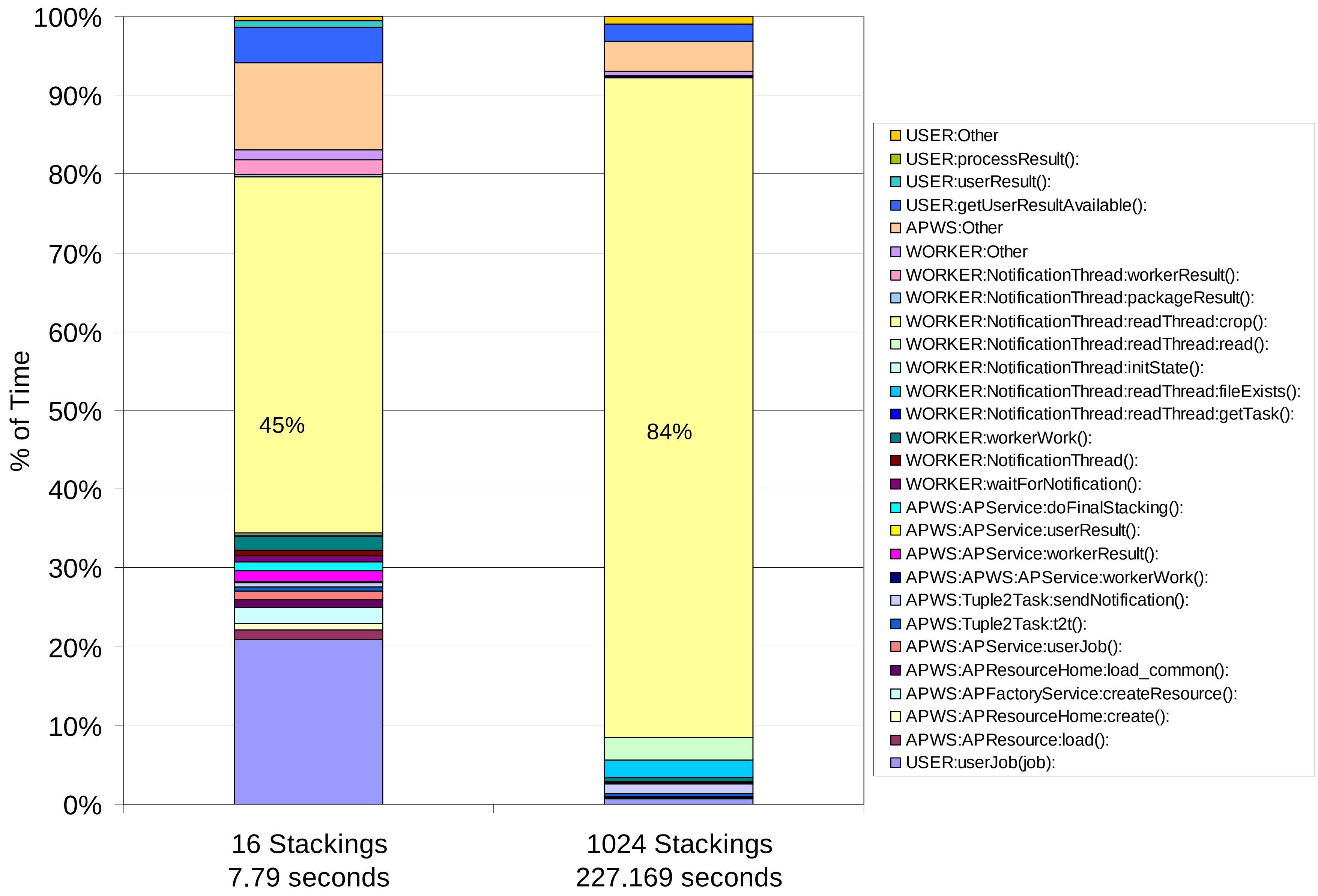
Stacking via the AstroPortal LAN GPFS in GZ Format



Stacking via the AstroPortal Local Disk in FIT Format

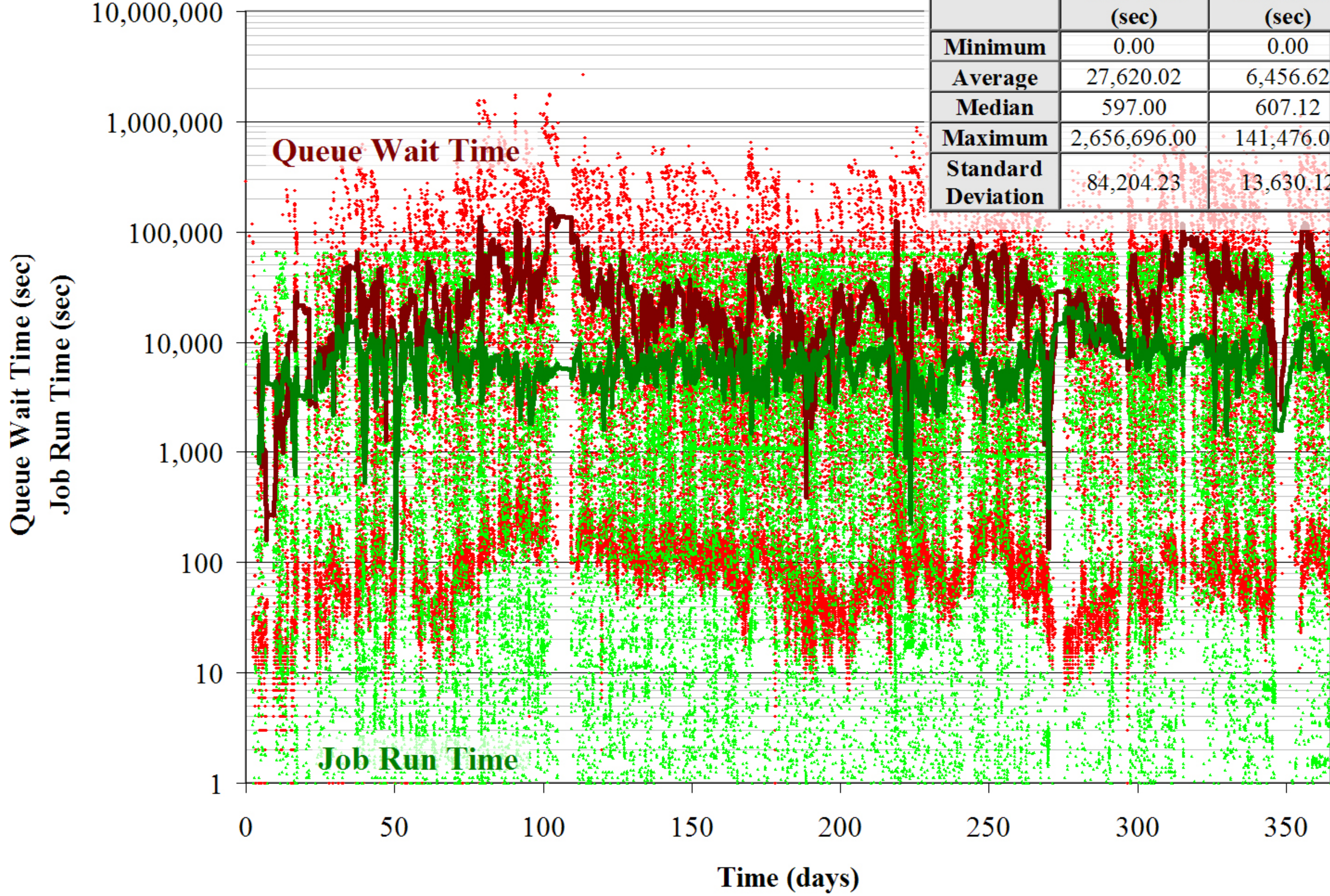


AstroPortal Stacking Profile LAN GPFS in GZ Format



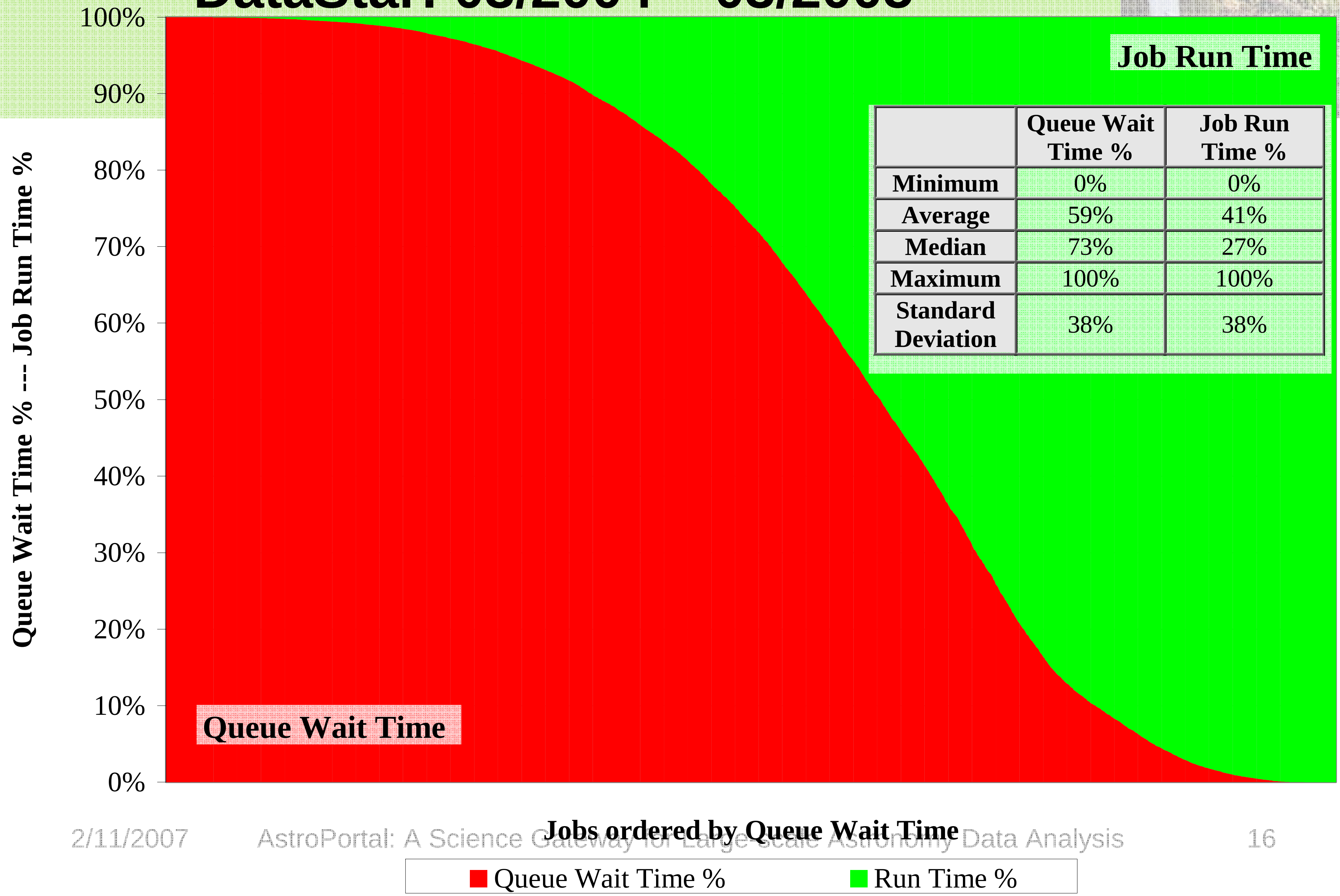
San Diego Supercomputer Center (SDSC) DataStar: 03/2004 – 03/2005

	Wait Time (sec)	Run Time (sec)
Minimum	0.00	0.00
Average	27,620.02	6,456.62
Median	597.00	607.12
Maximum	2,656,696.00	141,476.00
Standard Deviation	84,204.23	13,630.12



San Diego Supercomputer Center (SDSC)

DataStar: 03/2004 – 03/2005

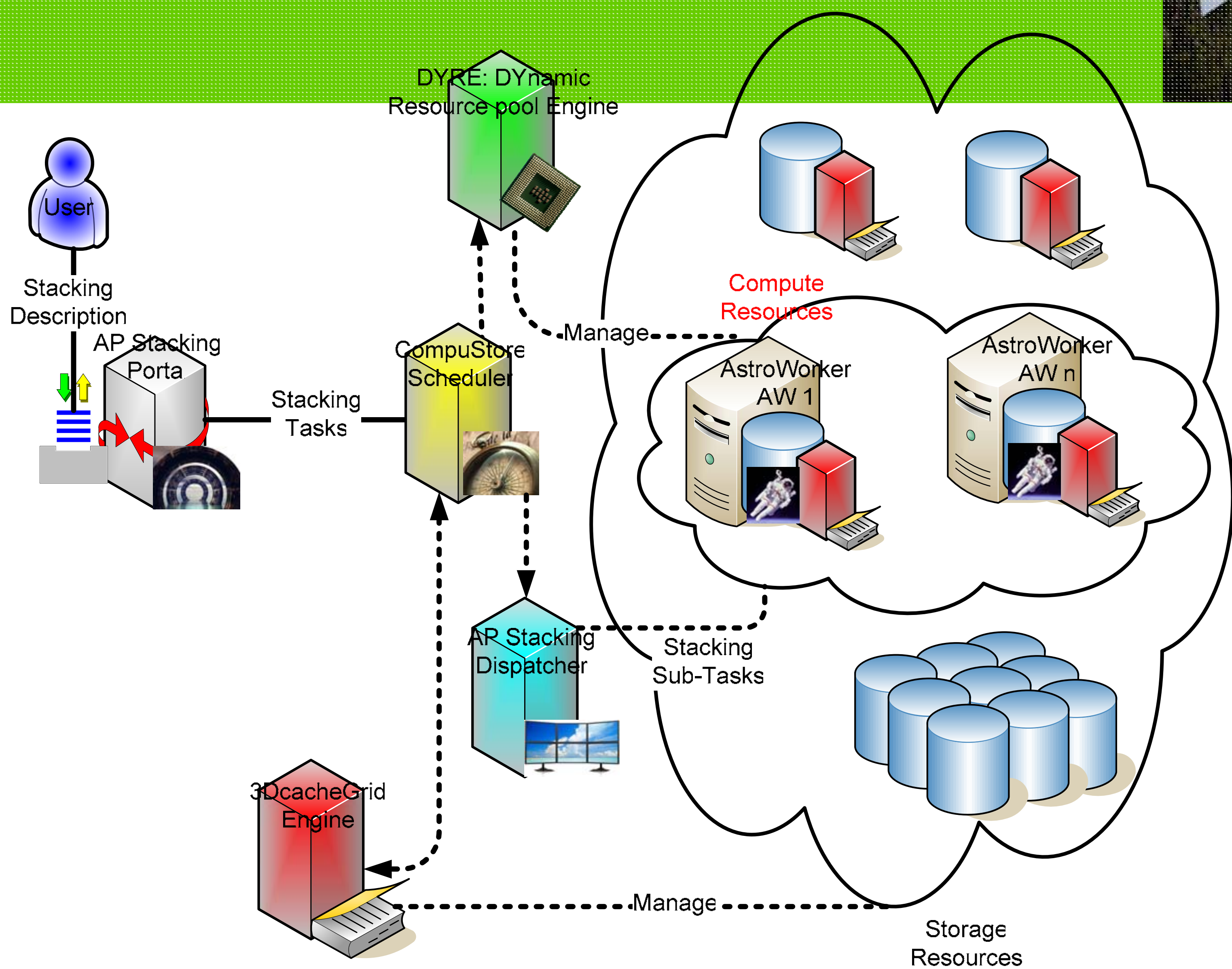


	Queue Wait Time %	Job Run Time %
Minimum	0%	0%
Average	59%	41%
Median	73%	27%
Maximum	100%	100%
Standard Deviation	38%	38%

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

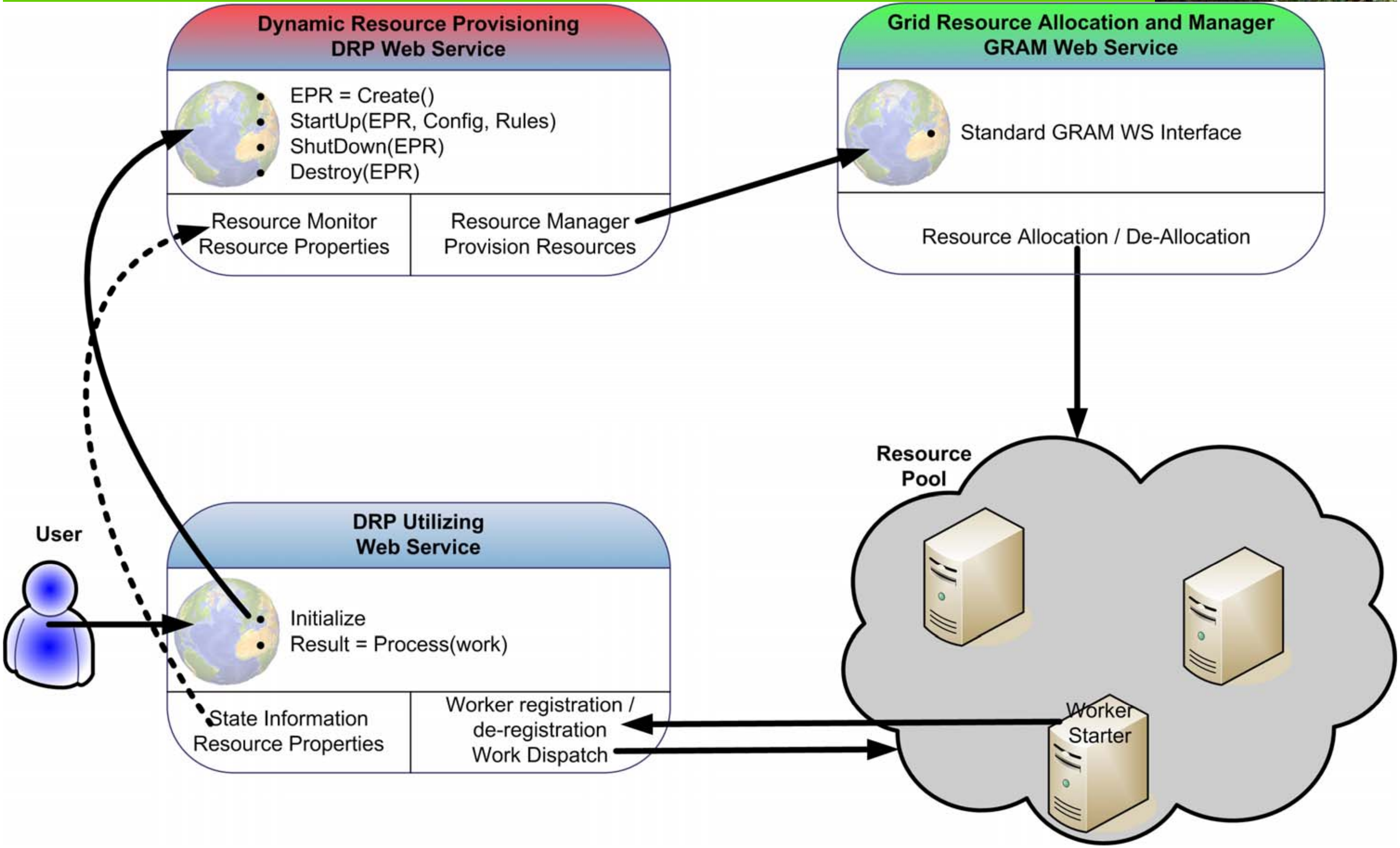
Proposed Opimizations



DRP: Dynamic Resource Provisioning

- State monitoring
- Resource allocation based on observed state
- Maintain a set of resources (even in the absence of lease extension mechanisms)
- Resource de-allocation based on observed state
- Exposes relevant information to other systems

DRP Architecture



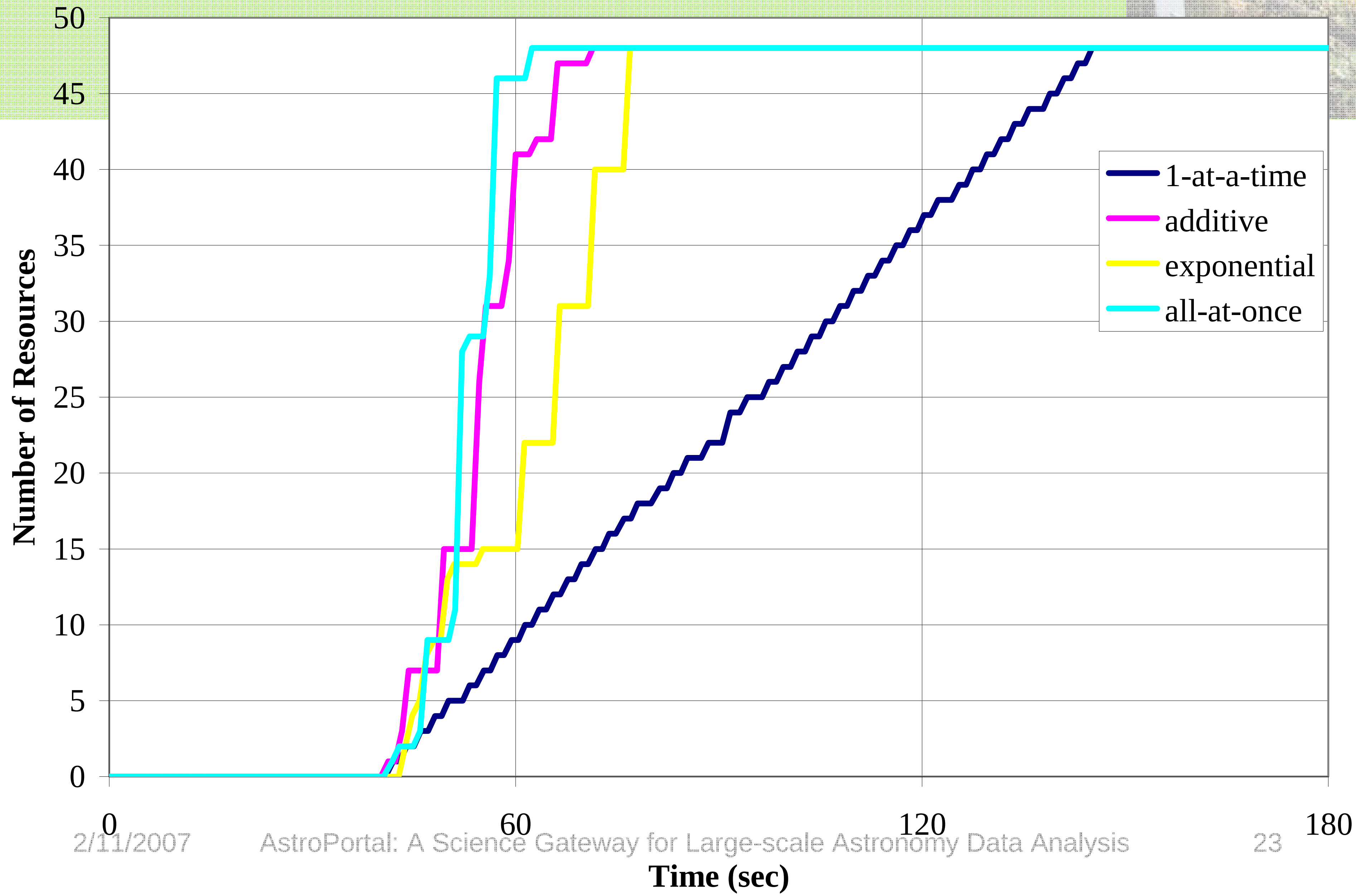
DRP Advantages

- 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

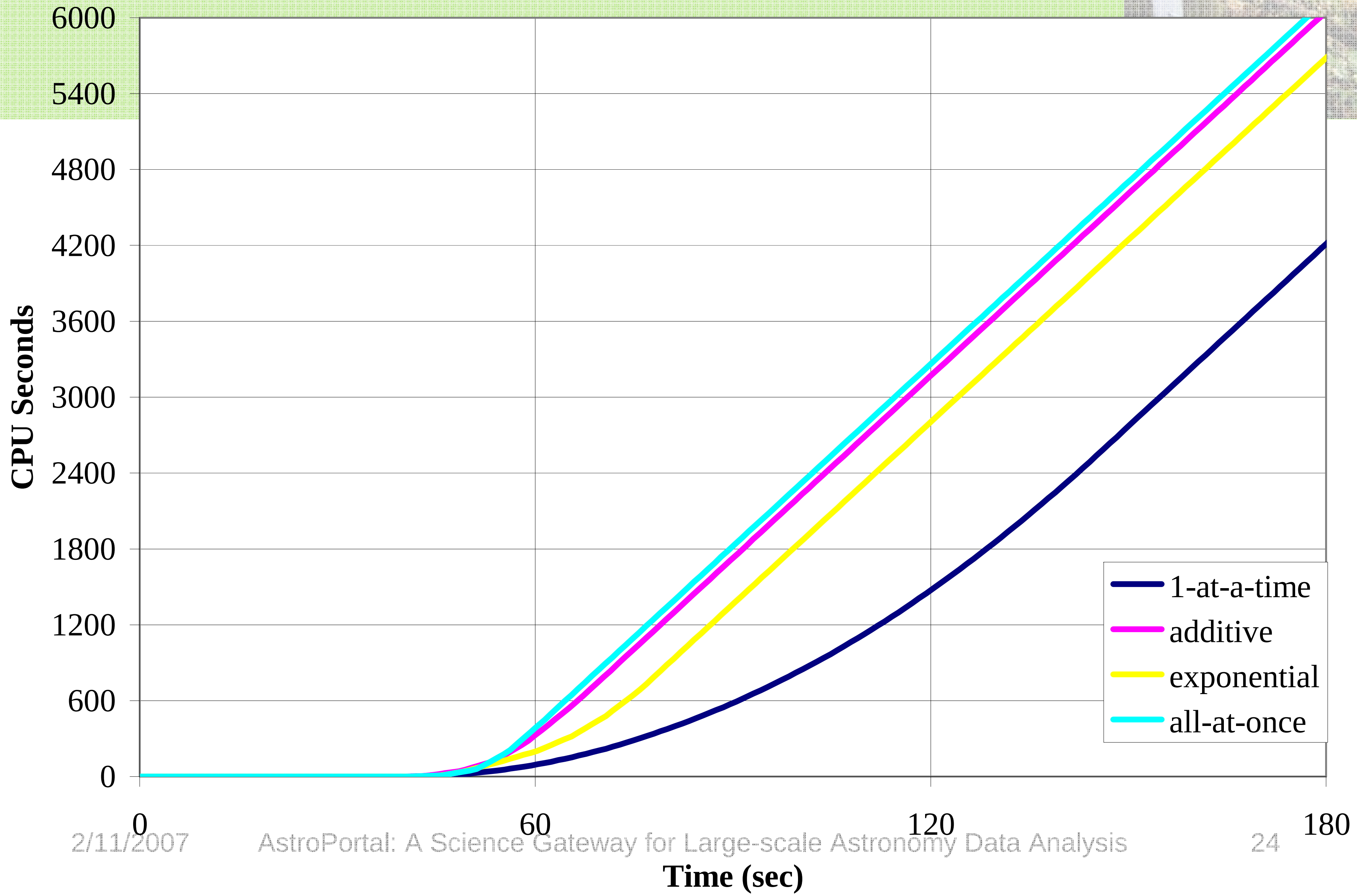
DRP Disadvantages

- 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

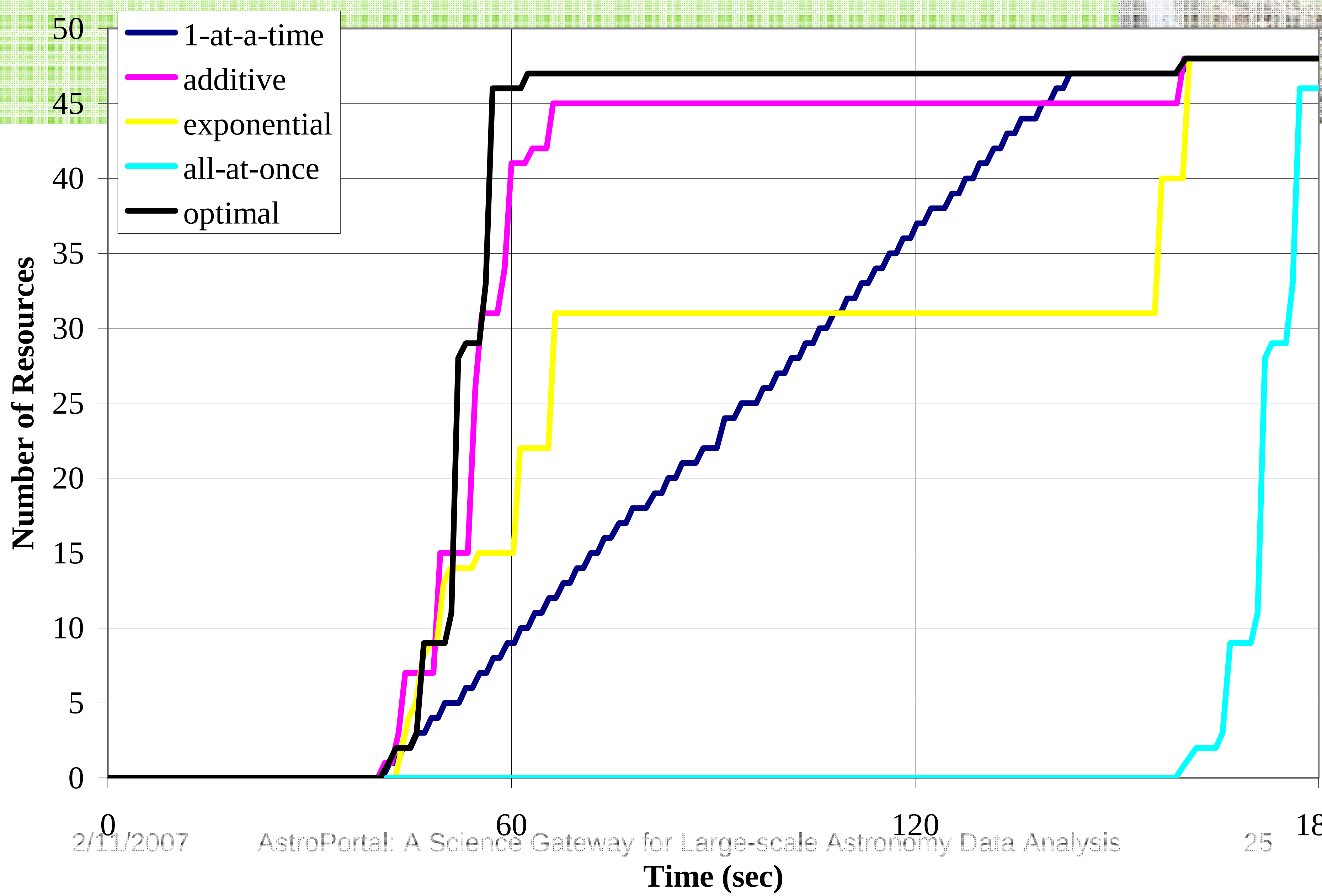
DRP Provisioning Latency



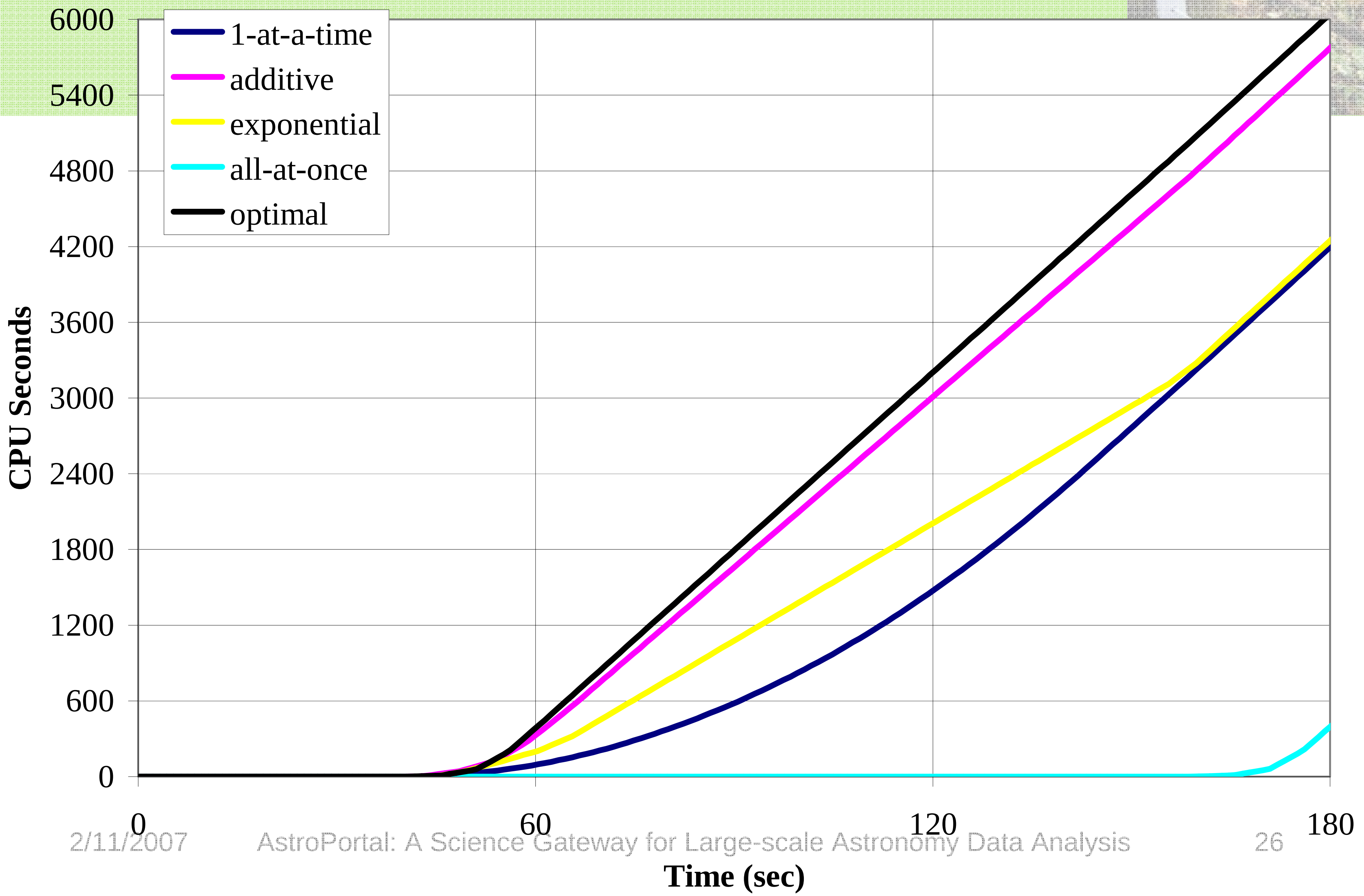
DRP Accumulated CPU Time



DRP Provisioning Latency



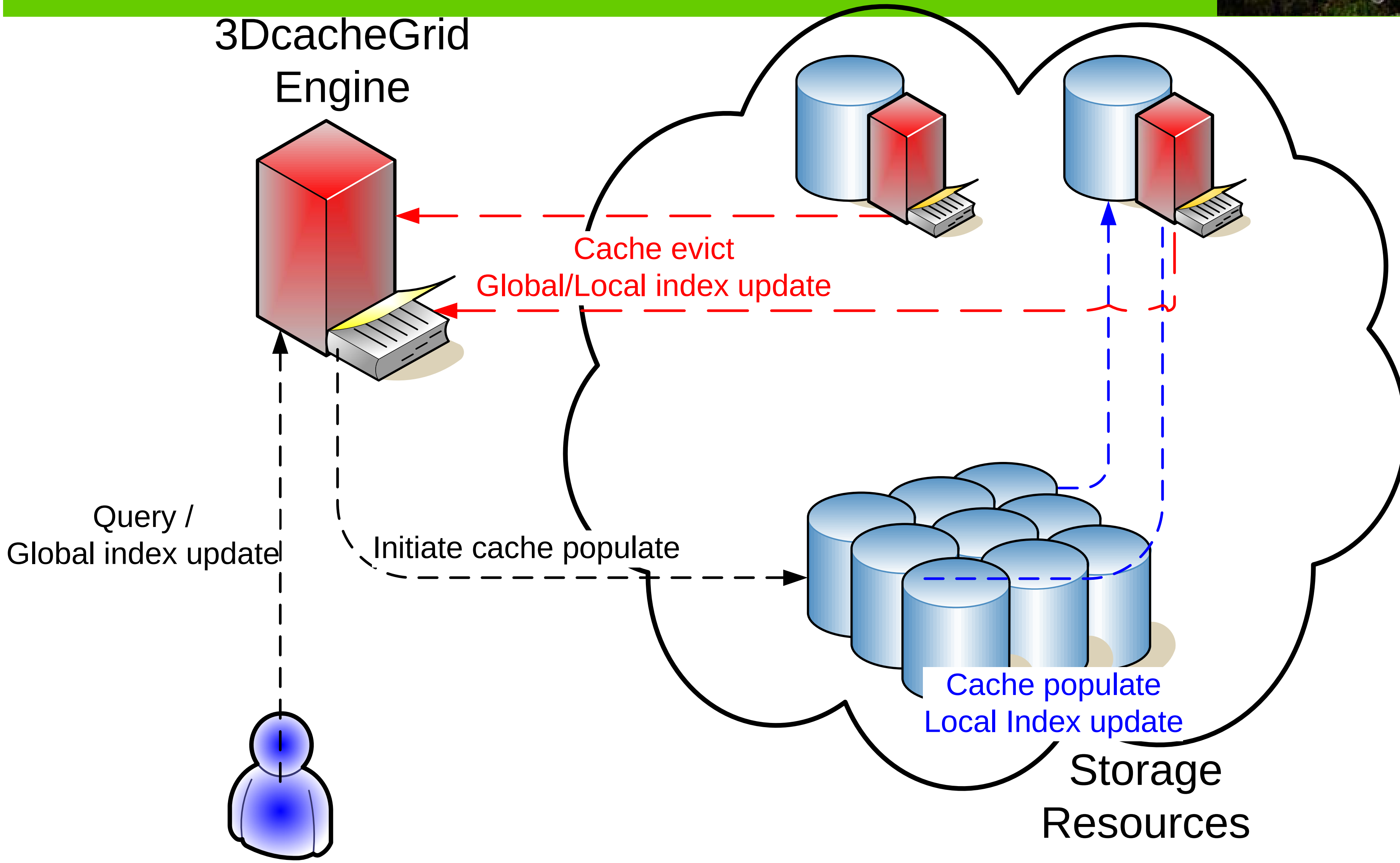
DRP Accumulated CPU Time



3DcacheGrid Engine: Dynamic Distributed Data cache for Grid Applications

- Performs data indexing necessary for efficient data discovery and access
- Cache eviction policy
  - RAND: Random
  - FIFO: First In First Out
  - LRU: Least Recently Used
  - Perfect LFU: Perfect Least Frequently Used
  - Hybrid Perfect LFU: Hybrid (using the object distribution in the dataset) Perfect Least Frequently Used
- Offers efficient management for large datasets along various dimentions
  - Number of files managed
  - Size of dataset
  - Number of storage resources used
  - Level of replication among the storage resources

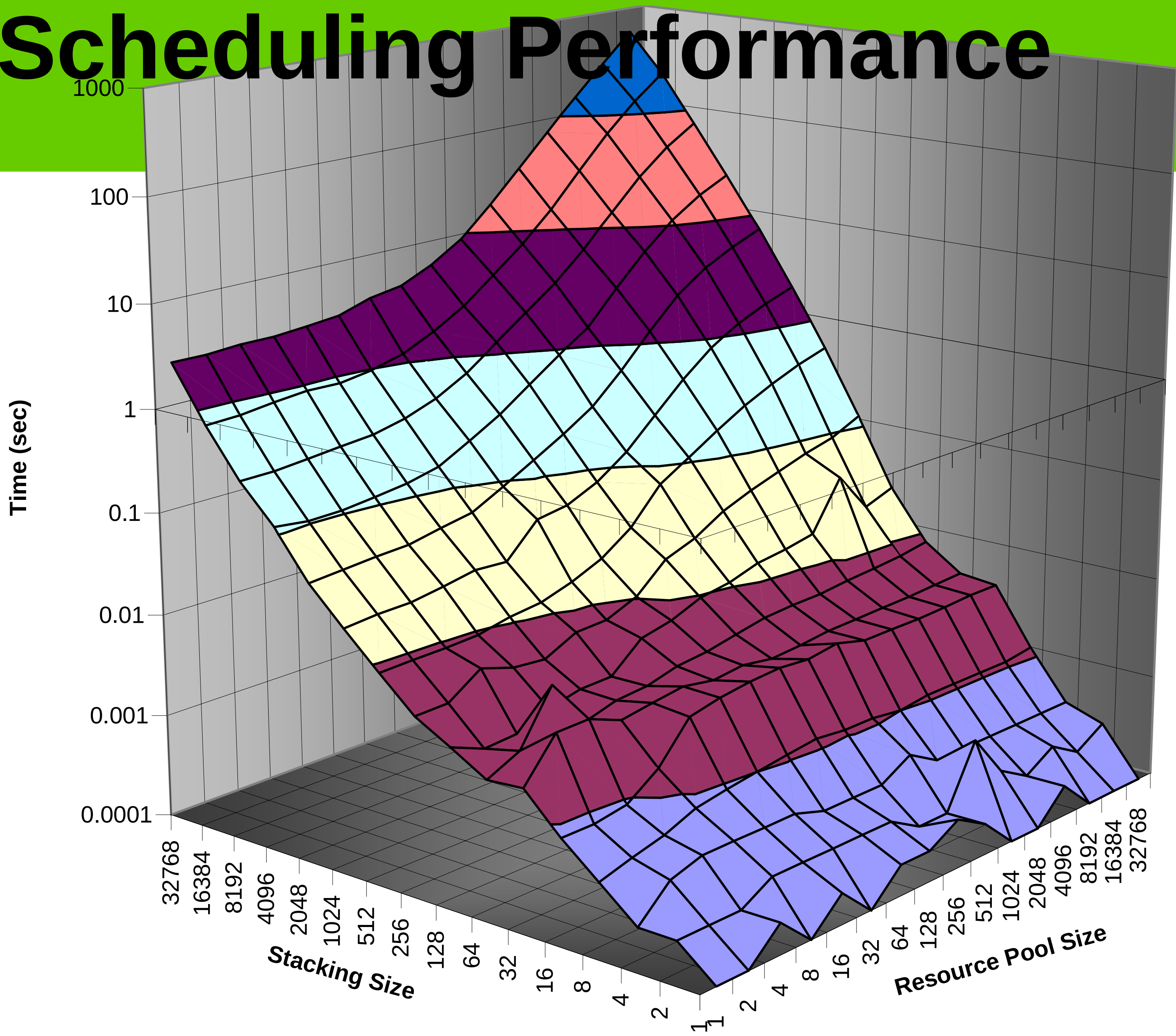
3DcacheGrid Architecture



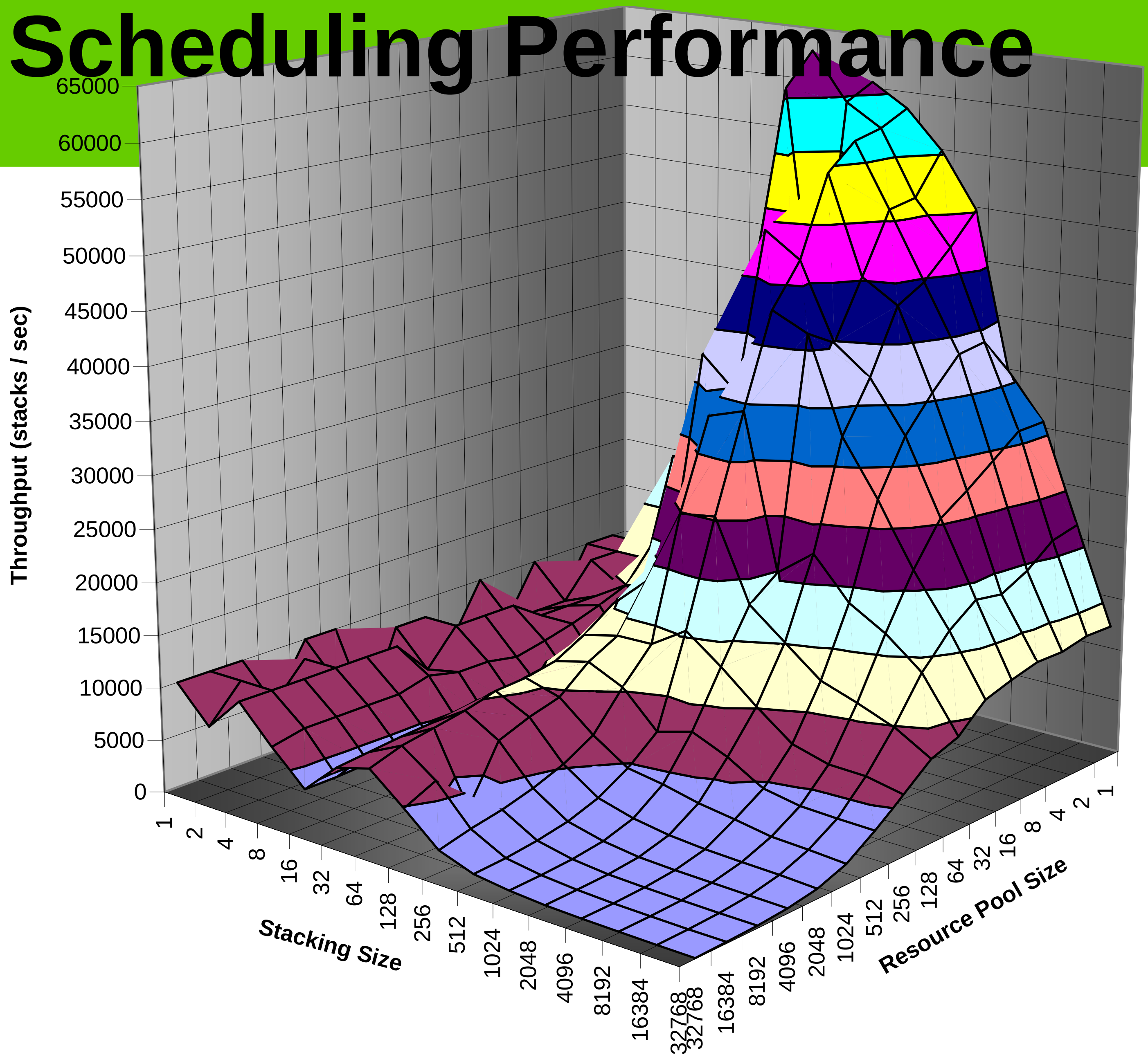
3Dcache Pros/Cons

- Pros:
  - Ease of application implementation: achieves a good separation of concerns between the application logic and the complicated data management task of large data sets
  - Improved performance with higher cache hits if data lcality is present
  - Improved scalability as the data I/O will be distributed over more resources with higher cache hits
  - Improved availability as cached data could be accessed without the need for the original data
  - Can enable compute scheduling to be data aware
- Cons:
  - Added complexity/overhead to a running system
  - Could produce worse overall performance than without 3DcacheGrid

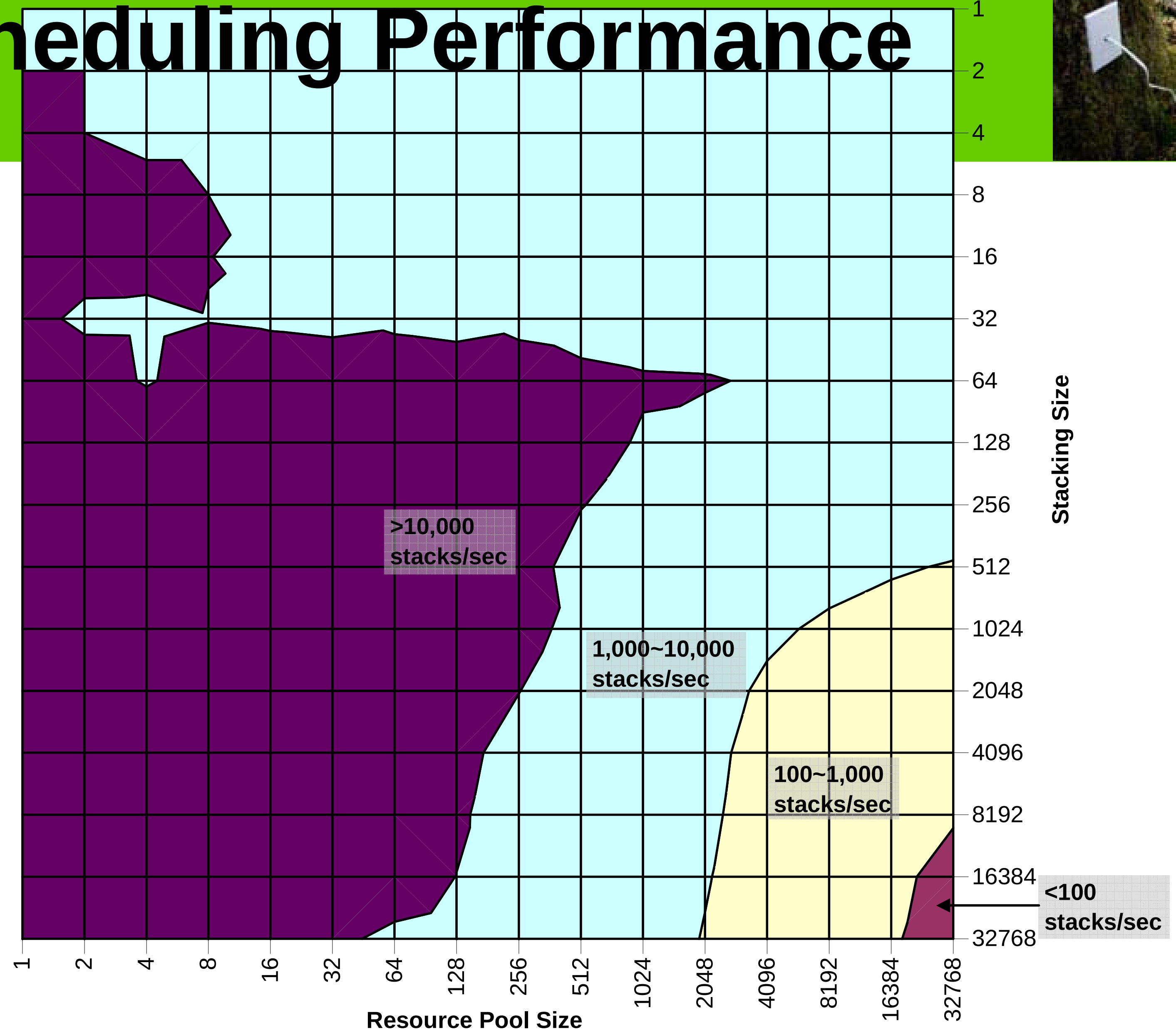
Data Management & Scheduling Performance



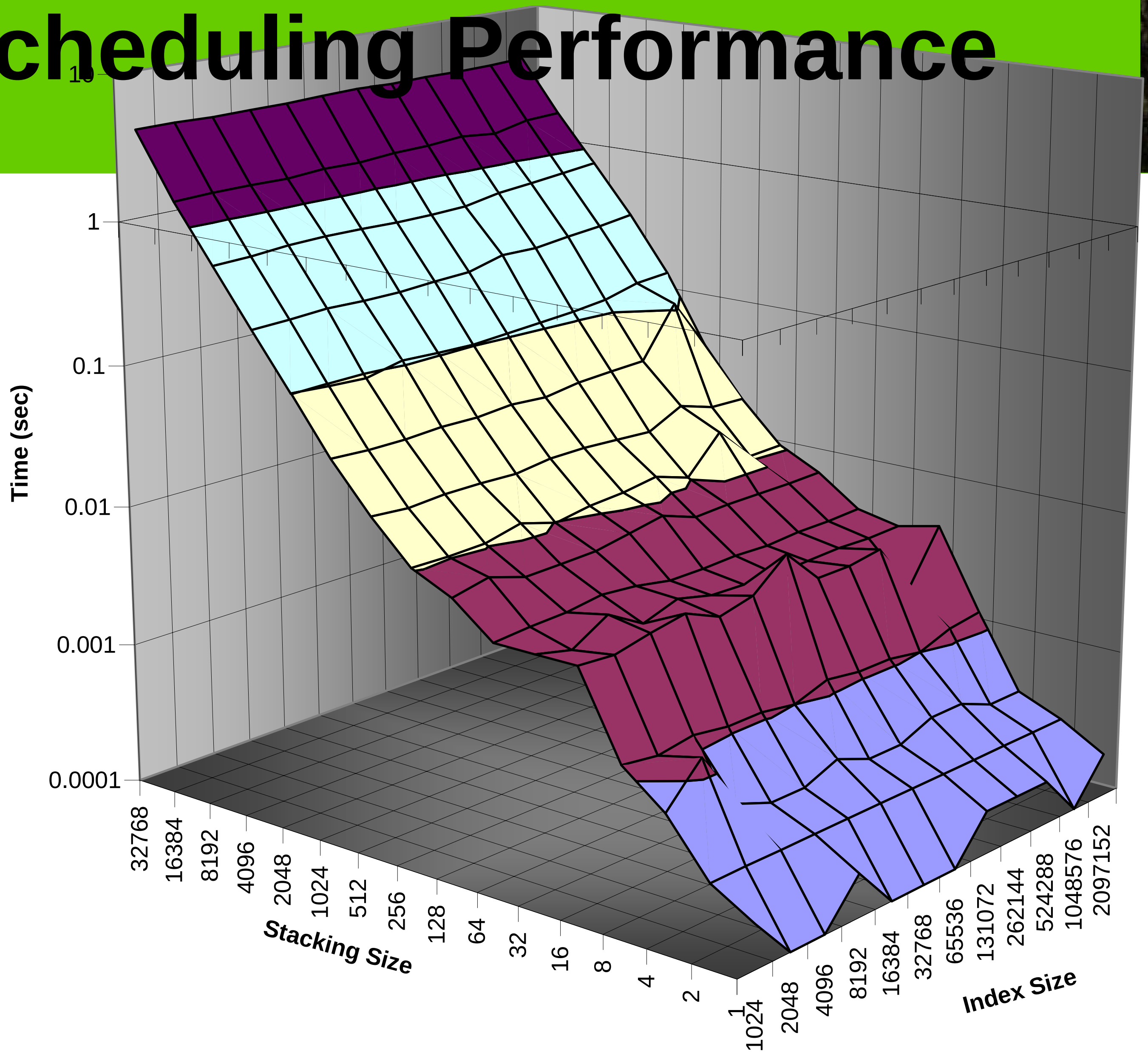
Data Management & Scheduling Performance



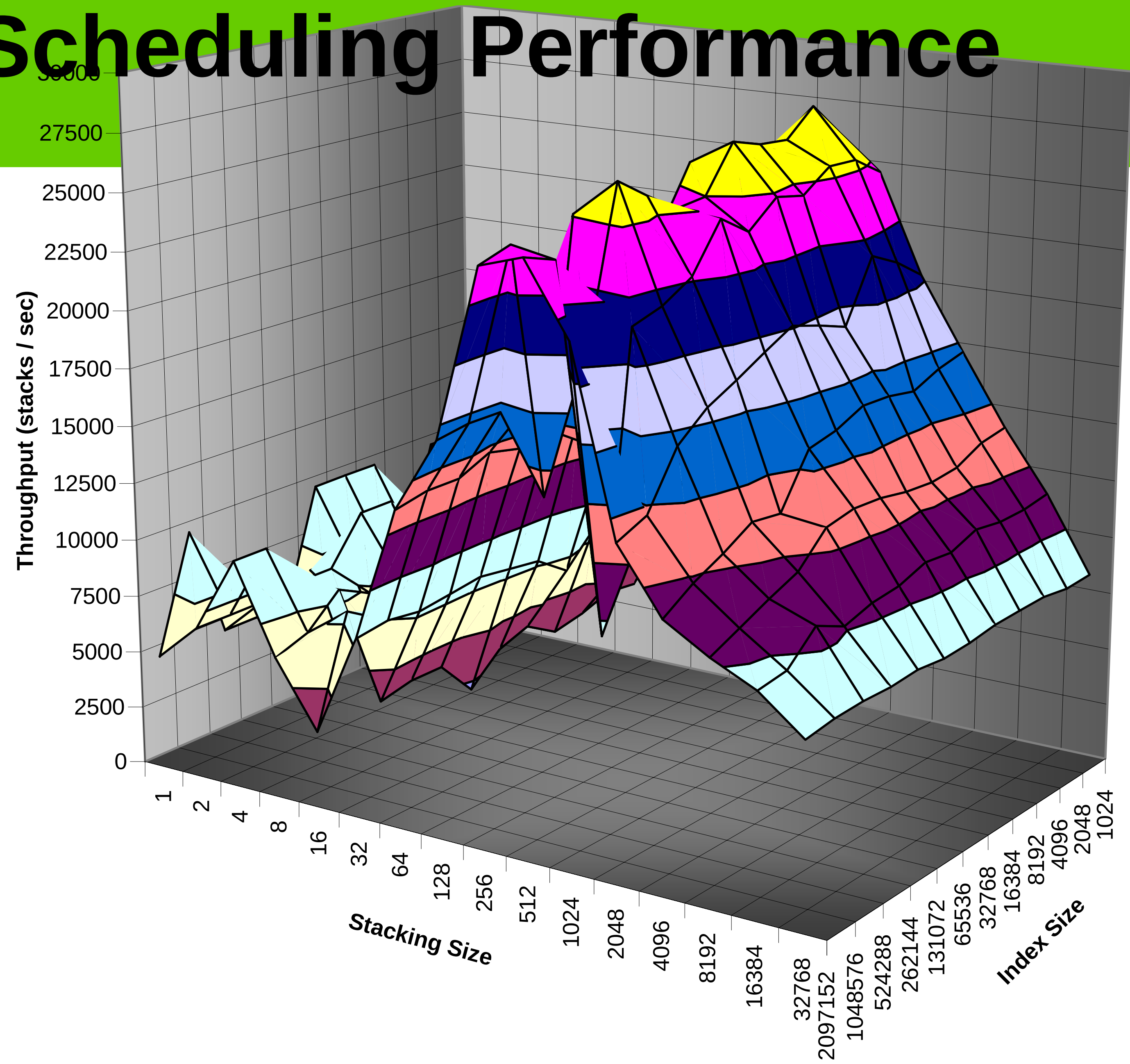
Data Management & Scheduling Performance



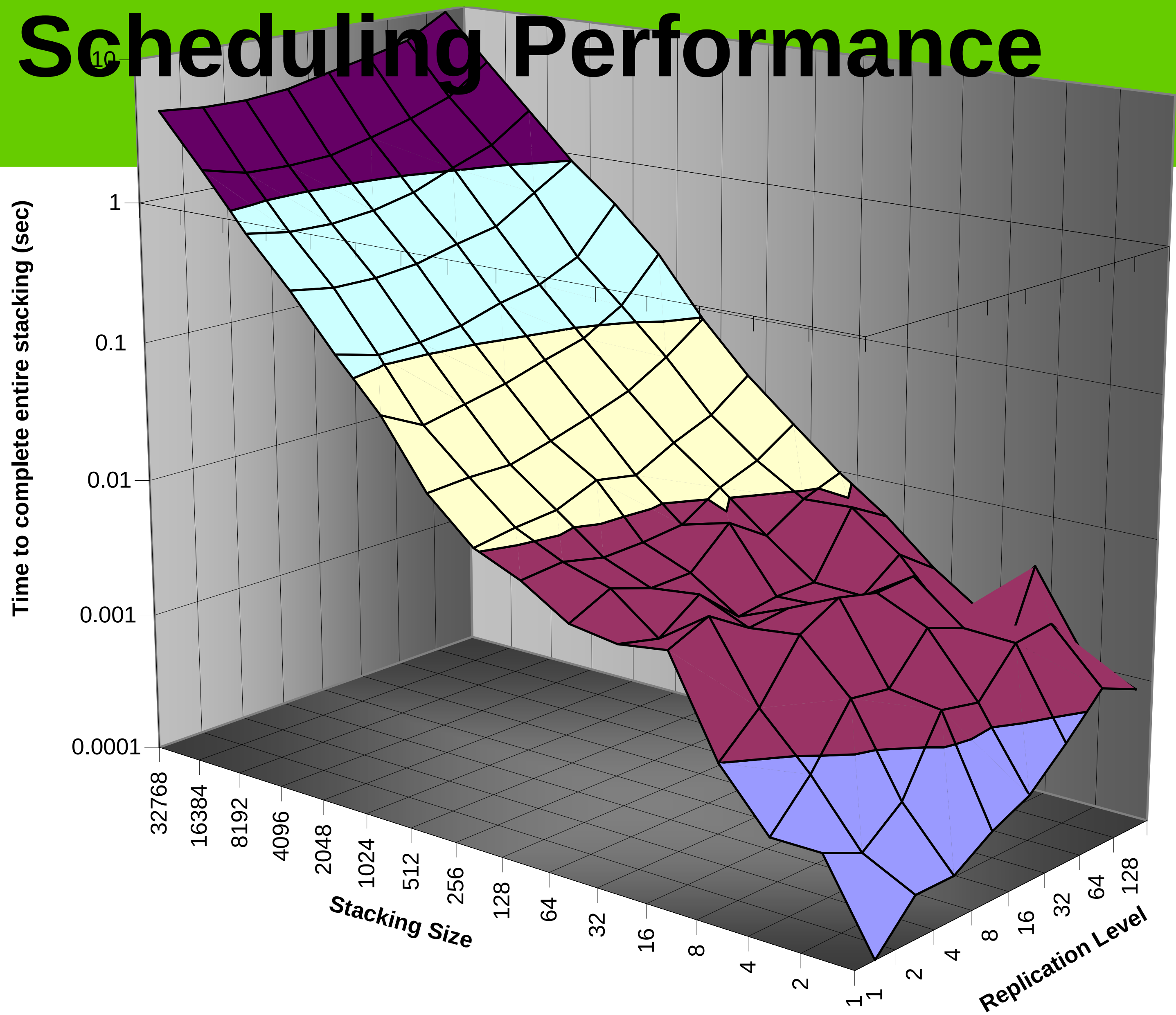
Data Management & Scheduling Performance



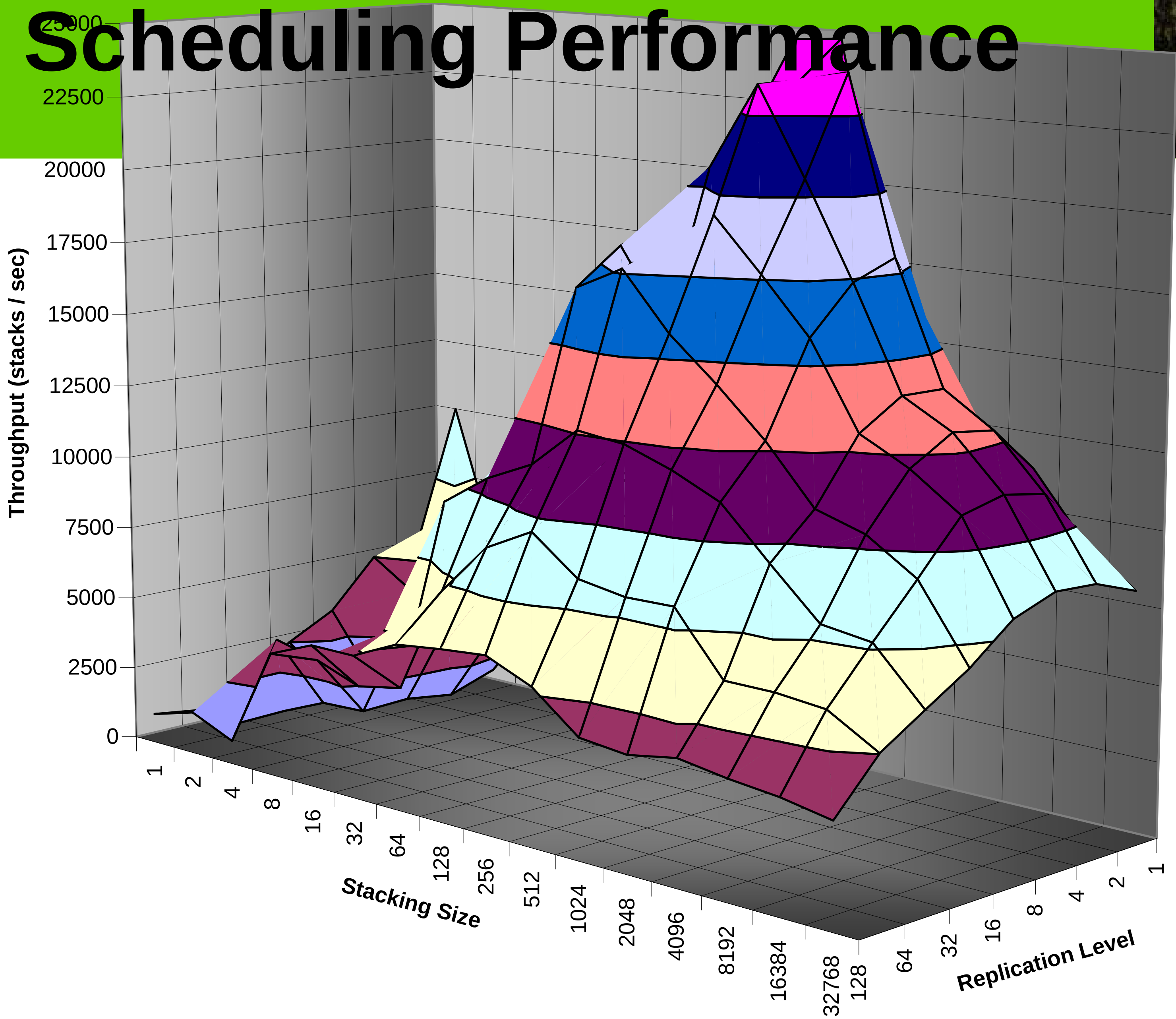
Data Management & Scheduling Performance



Data Management & Scheduling Performance



Data Management & Scheduling Performance



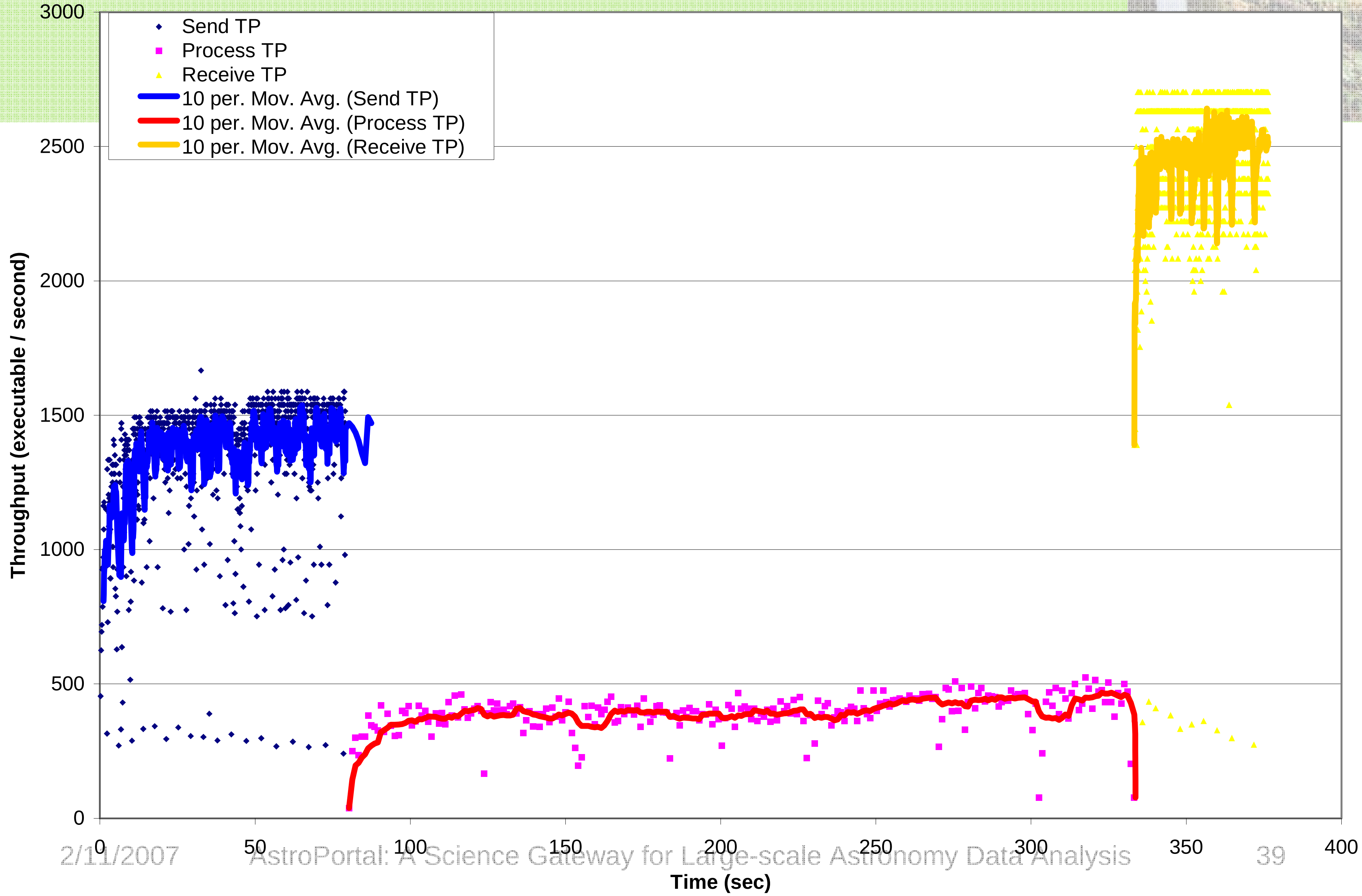
Data Management & Scheduling Performance Conclusions

- 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)

DeeF: Distributed execution environment Framework

- Binding glue connecting DRP, 3DcacheGrid, and CompuStore
- Allows the execution of aritrary code as well as pre-configured/installed code on remote resources managed by DRP
- Uses CompuStore to schedule tasks based on data locality of the caches
- Amortizes queue wait times over many tasks
- Enables the use of batch-scheduled Grids for interactive applications

DeeF Executing 100K Tasks



Questions?

- More information: http://people.cs.uchicago.edu/~iraicu/research/AstroPortal/
- AstroPortal Web Portal: http://s8.uchicago.edu:8080/AstroPortal/index.jsp
- 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.
  - Ioan Raicu. "***Harnessing Grid Resources to Enable the Dynamic Analysis of Large Astronomy Datasets***", NASA Ames Research Center GSRP Proposal for 10/2007 – 9/2008.
  - Ioan Raicu, Catalin Dumitrescu, Ian Foster. "***Dynamic Resource Provisioning in Grid Environments***", submitted to TeraGrid Conference 2007.
- Related papers that are in the writing pipeline (planning for SC07 and Grid07):
  - 3DcacheGrid: A Dynamic Distributed Data cache for Grid Applications
  - Data Aware Scheduling in High Throughput Computing
  - AMDASK: An Abstract Model for Data-Centric Task Farms
  - DeeF: A Distributed execution environment Framework
  - Enabling the Efficient Analysis of Large Astronomy Datasets with the AstroPortal version 2
  - Discoveries in the Sloan Digital Sky Survey Dataset using the "Stacking" Analysis Implemented by the AstroPortal