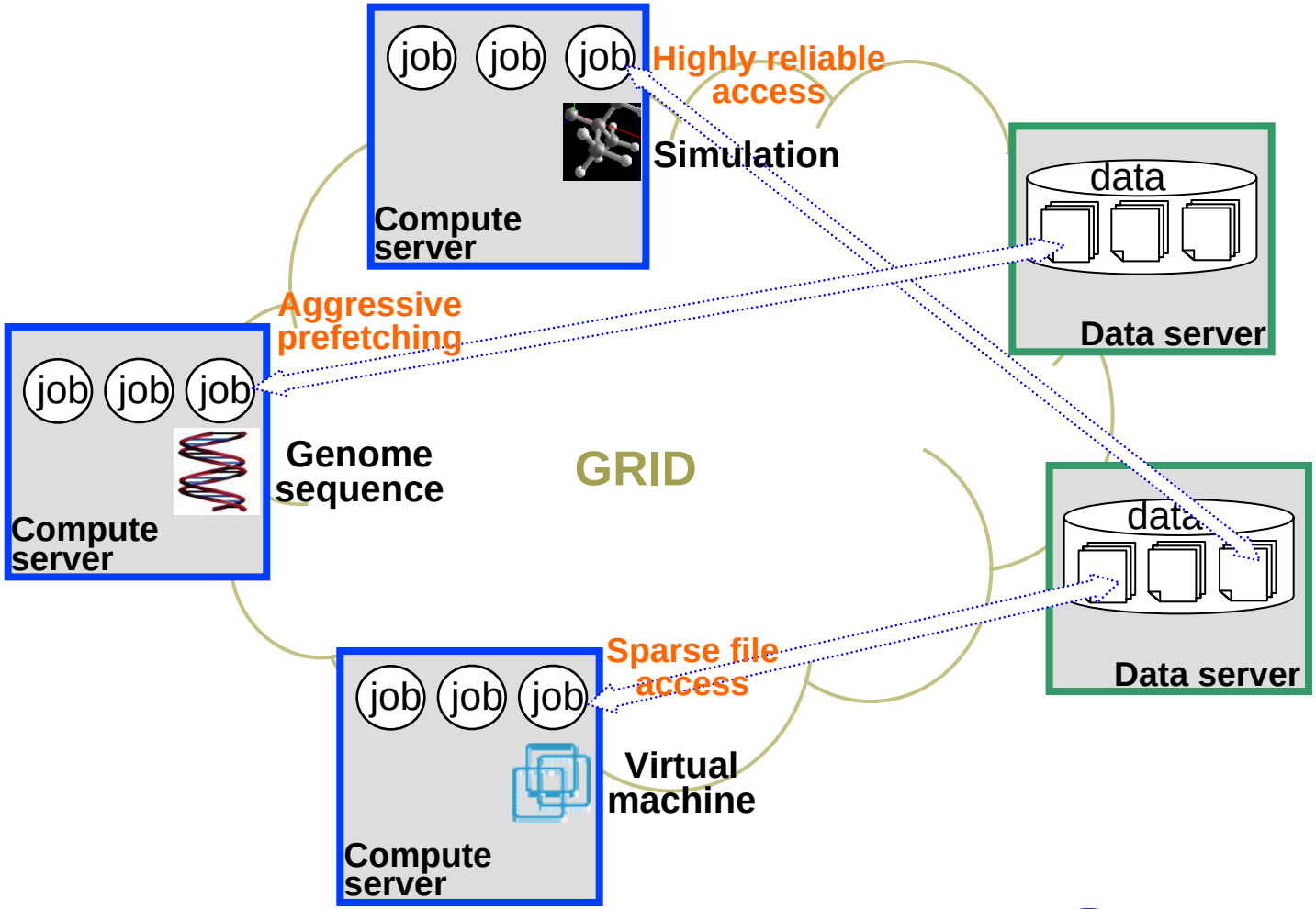
Towards Autonomic Grid Data Management with Virtualized Distributed File Systems

Ming Zhao, Jing Xu, Renato Figueiredo

Advanced Computing and Information Systems
Electrical and Computer Engineering
University of Florida
{ming, jxu, renato}@acis.ufl.edu

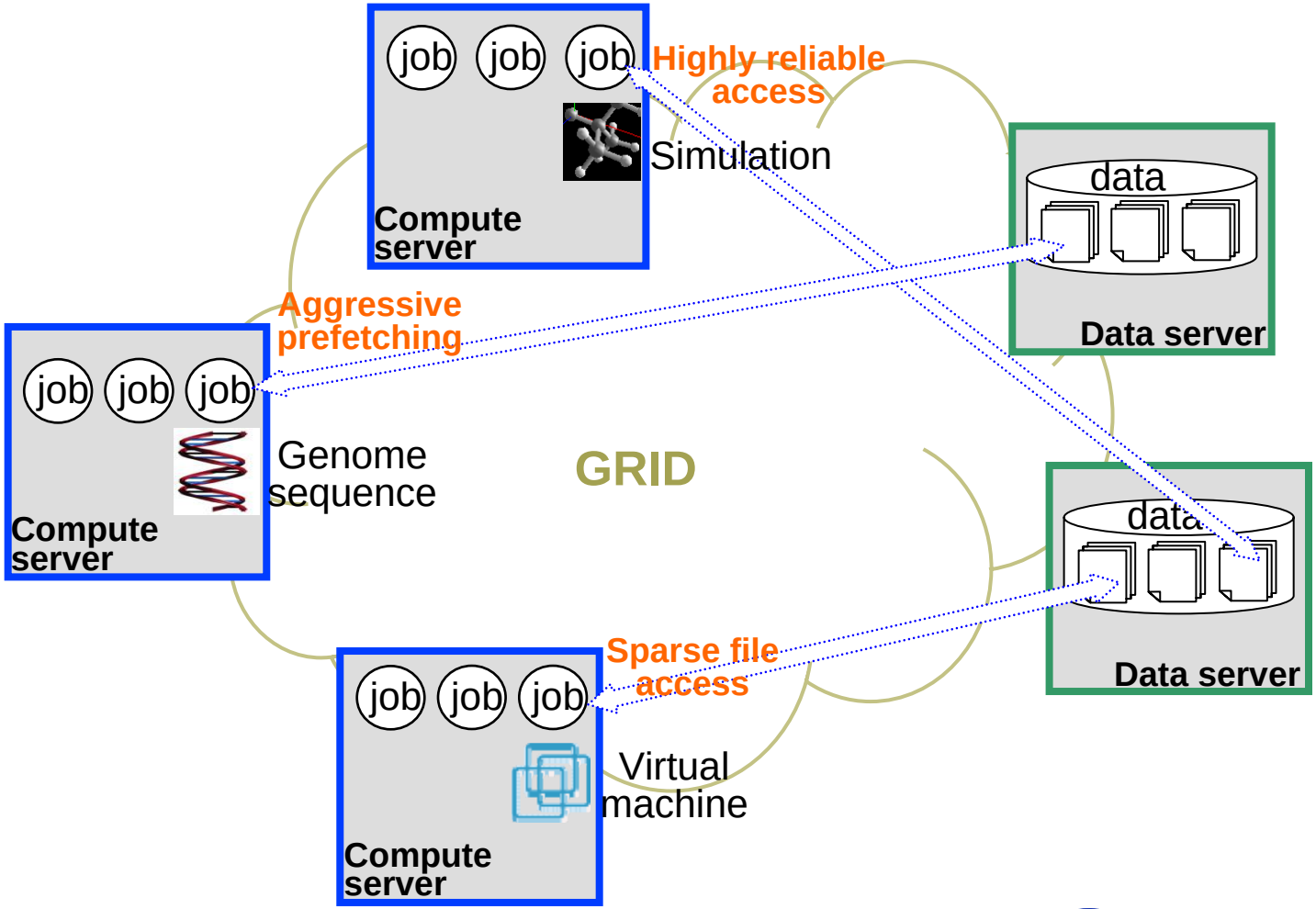
Motivation

- Grid data provisioning
  - Wide-area latency/bandwidth, dynamism of resources …
  - *How to provide data with application-tailored optimizations?*



Motivation

- Grid data management
  - Dynamic data access, resource sharing, changing resource availability …
  - *How to manage data provisioning in dynamically changing environments?*



Overview

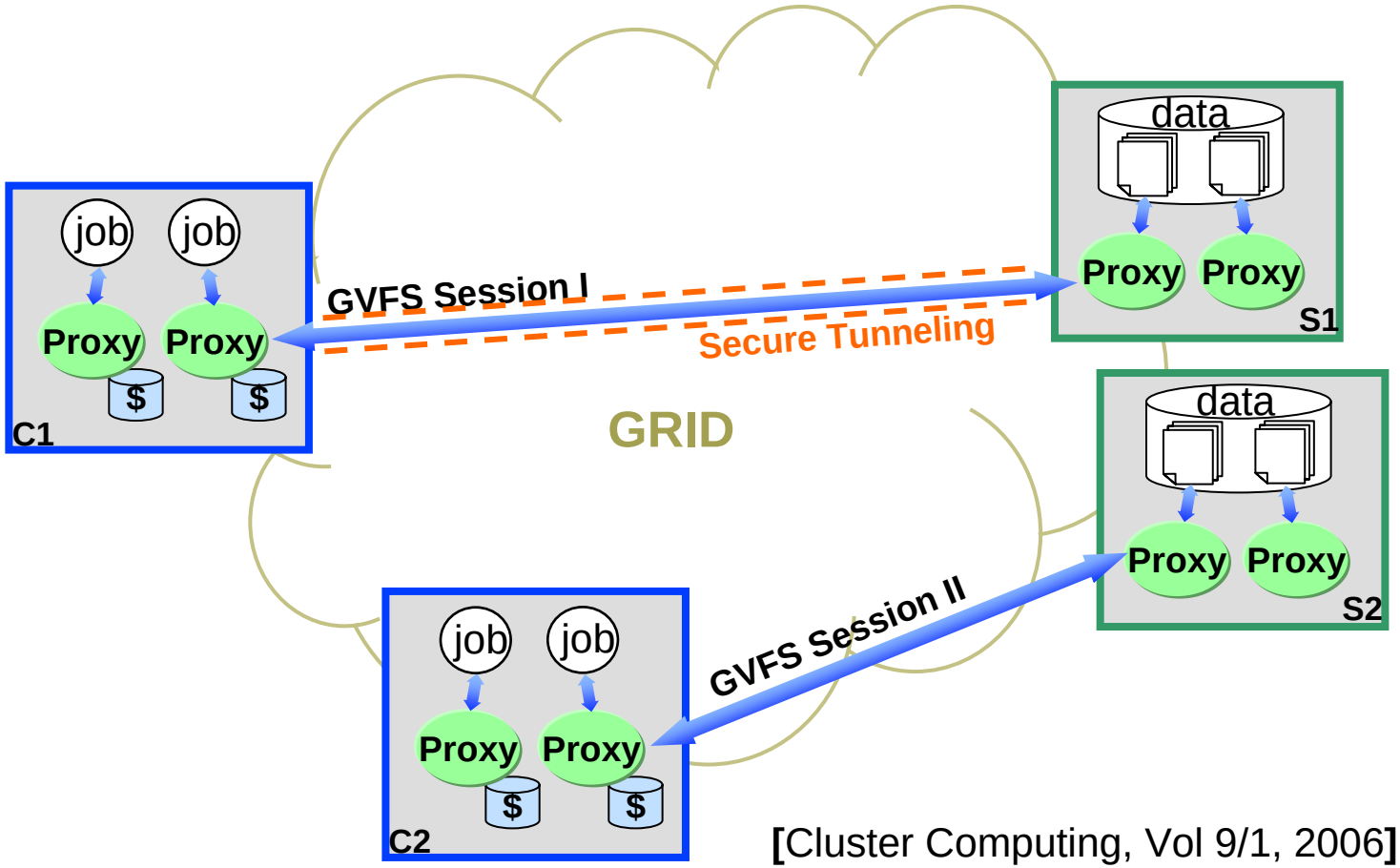
- Goal:
  - Efficient data management in heterogeneous, dynamic and large-scale Grid environments
- Challenges:
  - Application-tailored data provisioning
  - Data management in dynamically changing environment
- Contribution: Autonomic Grid data management
  - Virtualized Grid-wide file systems for user-transparent Grid data access with application-tailored enhancements
  - Autonomic data management services for self-managing, goal-driven control over Grid file system sessions

Outline

- Background
  - Grid virtual file system and data management
- Architecture
  - Autonomic data management services
- Evaluation
  - Experiments and analysis
- Summary
  - Related work, conclusion and future work

Grid Virtual File Systems

- **G**rid **V**irtual **F**ile **S**ystem (GVFS)
  - Distributed file system virtualization through user-level NFS proxies
  - Enhancements for Grid environment (disk caching, secure tunneling)
  - Dynamic, independent, application-tailored GVFS sessions



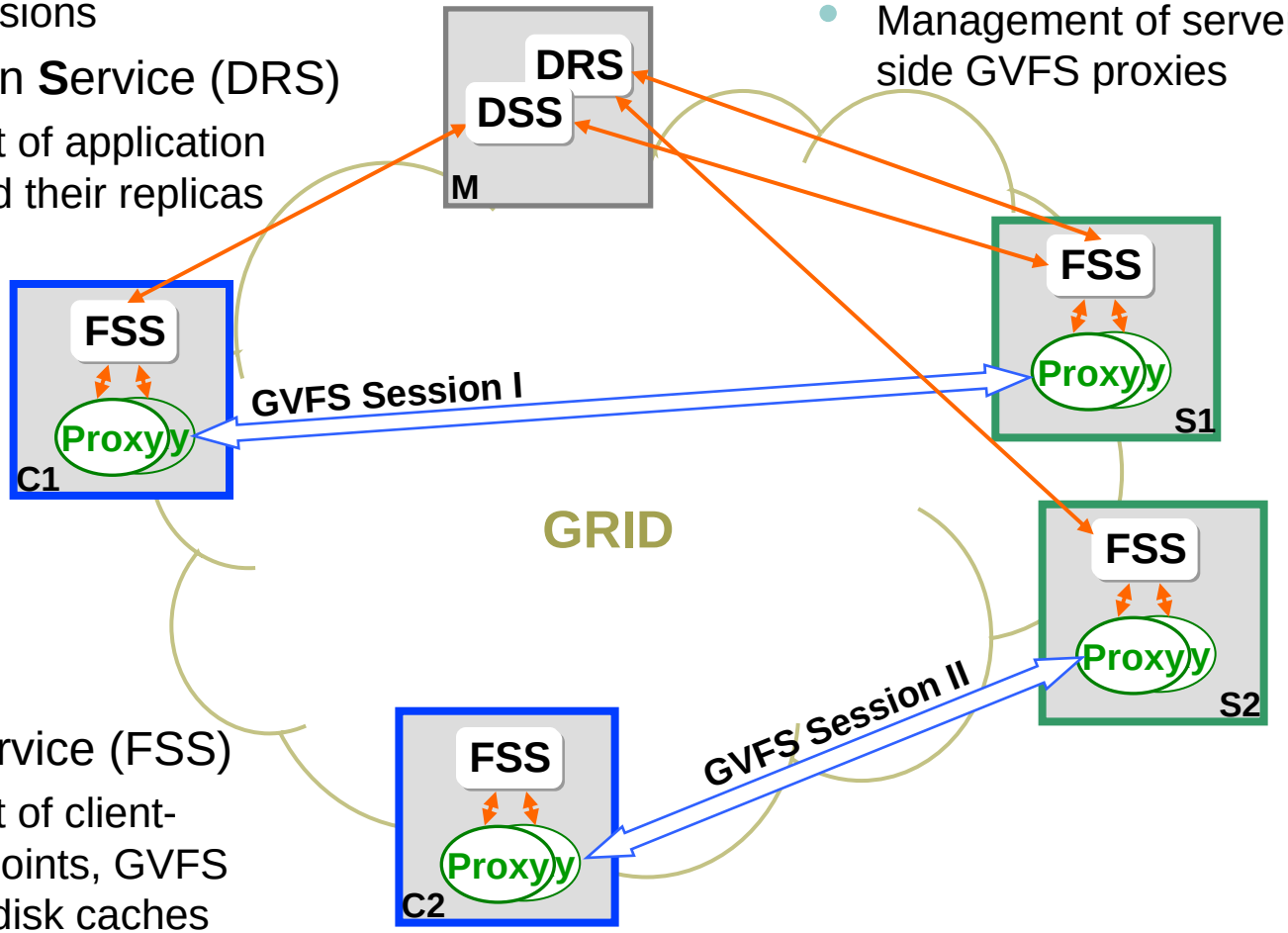
[Cluster Computing, Vol 9/1, 2006]

Data Management Services

- WSRF-based management services for GVFS sessions

- **D**ata **S**cheduler **S**ervice (DSS)
  - Scheduling and customization of GVFS sessions
- **D**ata **R**eplication **S**ervice (DRS)
  - Management of application data sets and their replicas

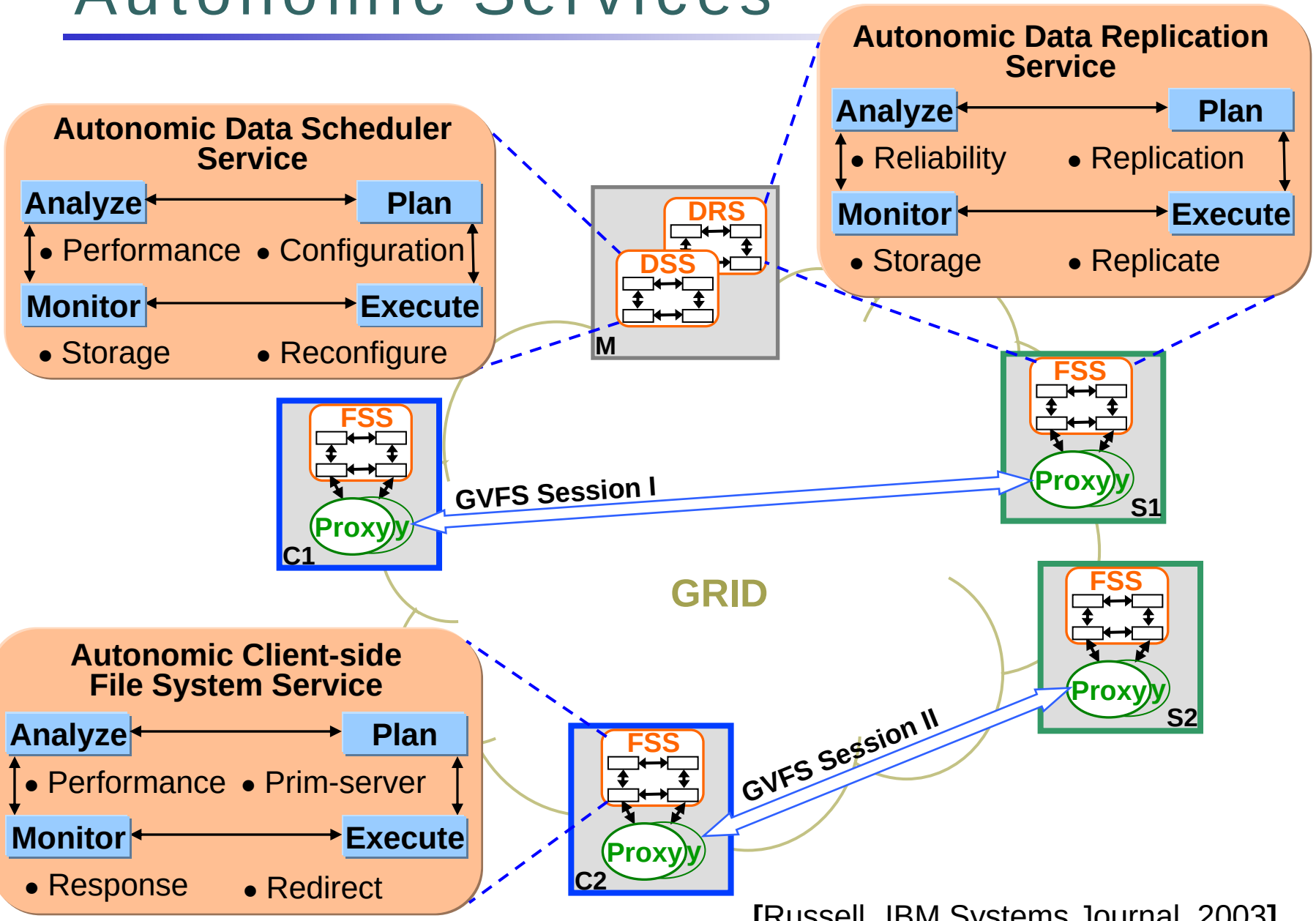
- Server-side **F**ile **S**ystem **S**ervice (FSS)
  - Management of server-side GVFS proxies



- Client-side **F**ile **S**ystem **S**ervice (FSS)
  - Management of client-side mount points, GVFS proxies and disk caches

[HPDC'2005]

Autonomic Services

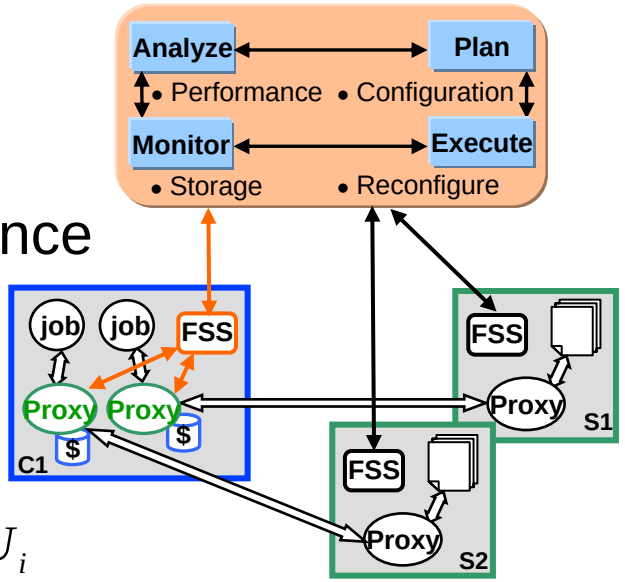


[Russell, IBM Systems Journal, 2003]

Autonomic Data Scheduler Service

- Manages concurrent GVFS sessions on shared resources
  - Considers both application performance and resource utilization policy
  - Session *i*'s utility:

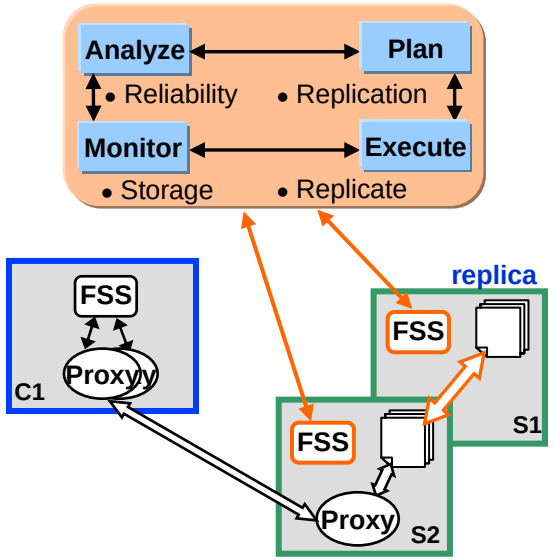
    $U_i = Performance_i * Priority_i$
  - Configures sessions to maximize $\sum_i U_i$



- E.g. Disk cache management
  - Hit rate is important in wide-area environment
  - Allocates client-side storage among sessions
    - Monitors storage usage via client-side FSS
    - Configures the sessions' caches to maximize global utility
    - Applies configurations via client-side FSS

Autonomic Data Replication Service

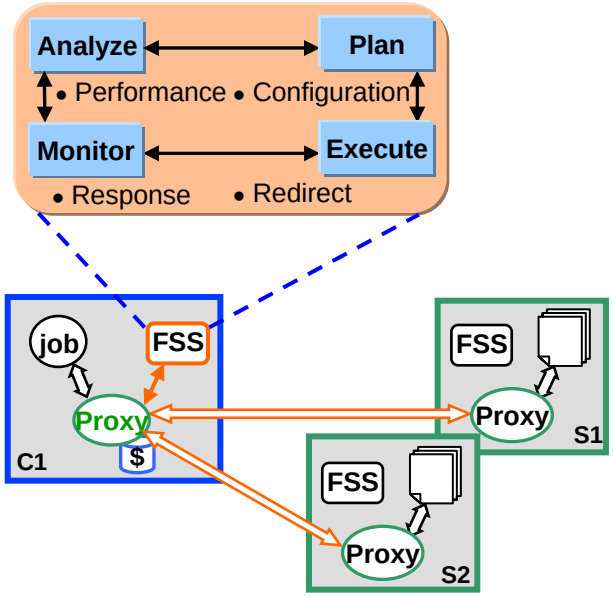
- Replication degree and placement
  - Increases reliability against server-side failures (crash, network partitioning)
    - The reliability of a session's data set d : $Reliability_d > R_{min}$
  - Reduces replication overhead
    - The cost of creating replicas for data set d $Cost_d < C_{max}$
  - Replaces replicas based on utility
    - $U_d = Value_d * Reliability_d$, maximizes $\sum_i U_d$



- Replica regeneration
  - Monitors server status via server-side FSS
  - Reconfigures sessions' replicas after a failure
  - Generates replicas via server-side FSSs

Autonomic File System Service

- Fail-over against server failures
  - Minimizes impact on the application
  - Non-interrupt session redirection
    - Monitors server response time
    - Detects failure when request times out
    - Redirects session to backup server
    - Regenerate replicas via DSS/DRS



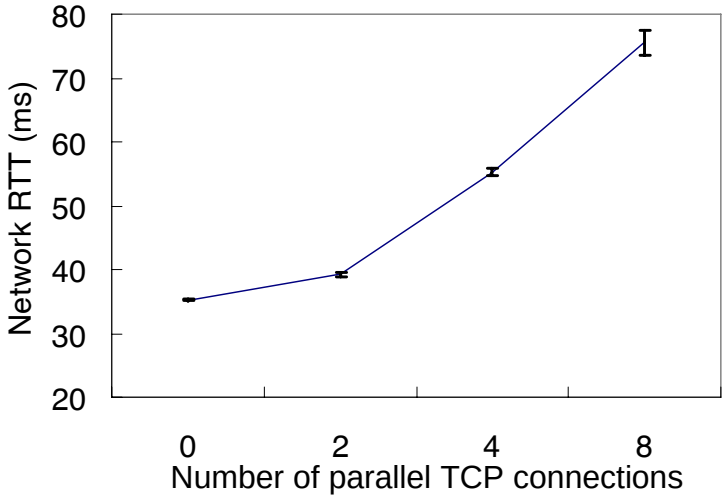
- Primary server selection
  - Maximizes performance against dynamic environment
    - Monitors connections with effective, low-overhead mechanism: Small random writes to a hidden file on the mounted partition
    - Predicts performance with simple effective forecasting algorithm
    - Chooses the best connection and switches transparently

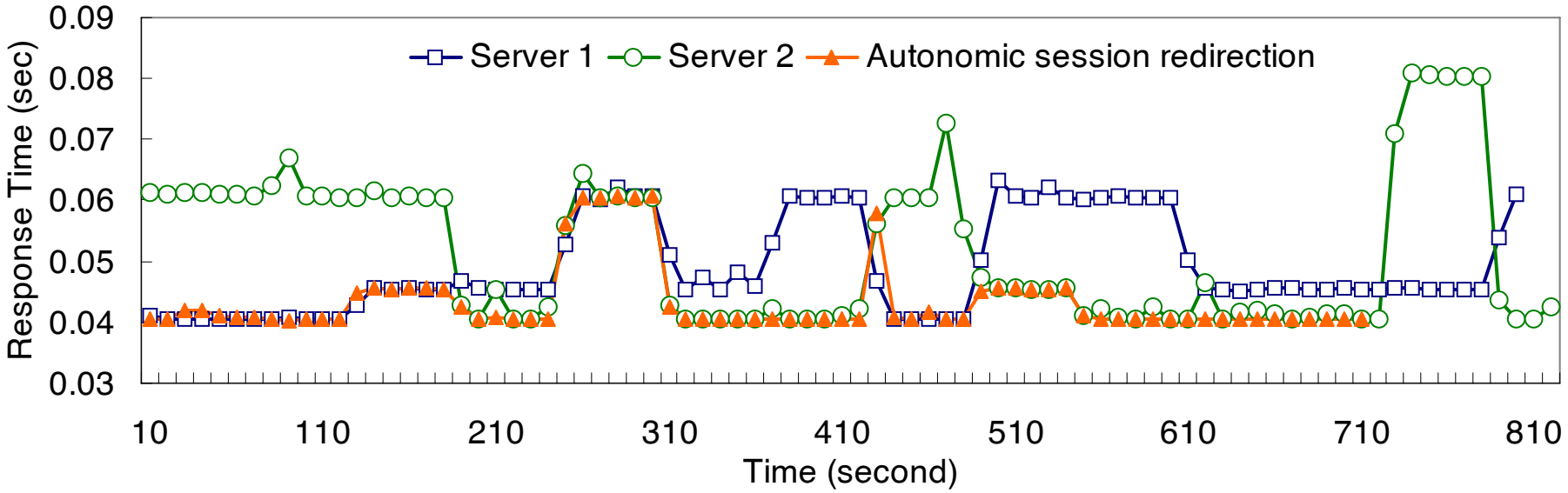
Experimental Setup

- File system clients and servers
  - VMware-based virtual machines
- Wide-area networks
  - NIST Net emulated links
- I/O part of Grid applications
  - IOzone file system benchmark
- Grid data management
  - User-level NFS v3 based GVFS proxies
  - WSRF-Lite based management services

Autonomic Session Redirection

- Scenario (I)
  - Data transfer under ***network fluctuation***
  - E.g. caused by parallel TCP transfers
- Setup
  - Client connected to two servers with independently emulated WAN links
  - Randomly applied latencies with values from a real wide-area measurement

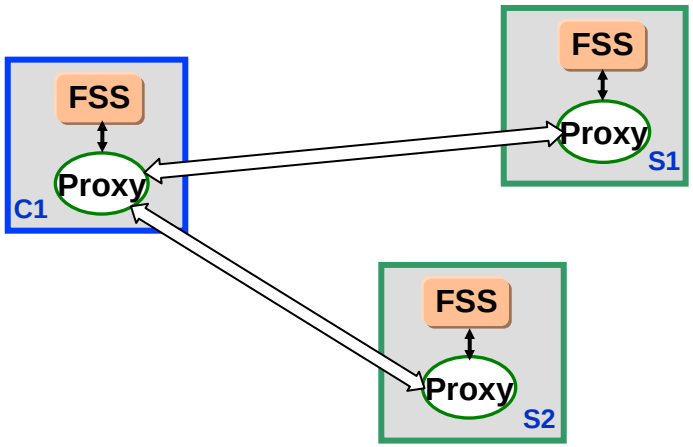


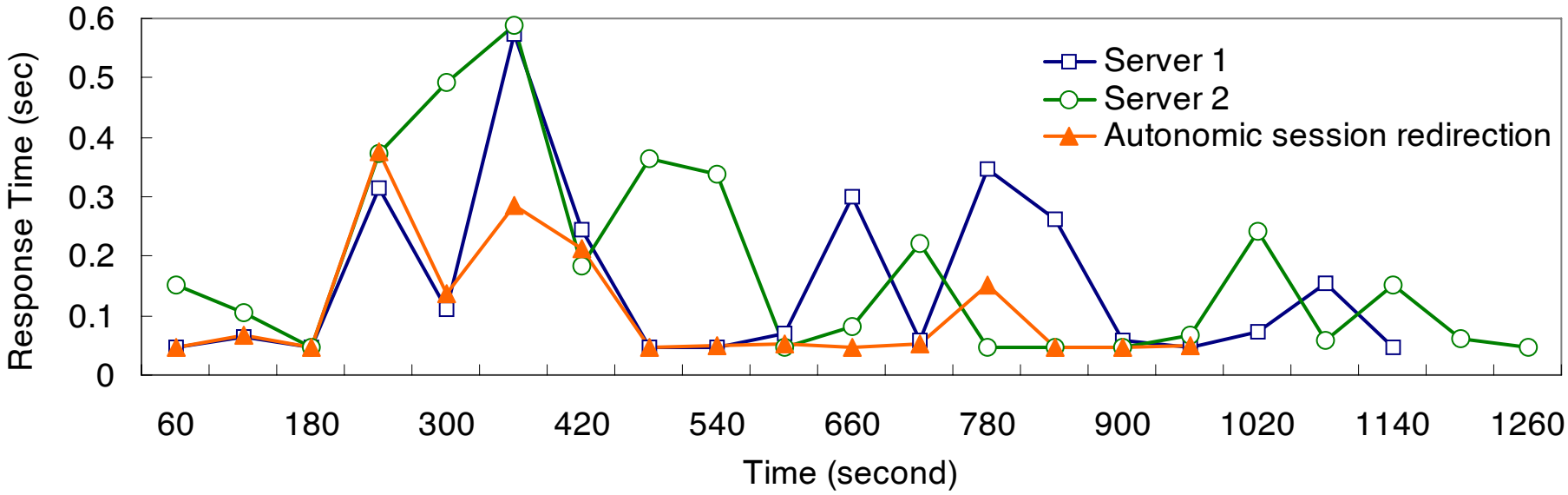


- **Autonomic session redirection achieves the best performance by adapting to the changing network condition**

Autonomic Session Redirection

- Scenario (II)
  - Data transfer under ***server load variation***
- Setup
  - Randomly generated background load applied on the data servers

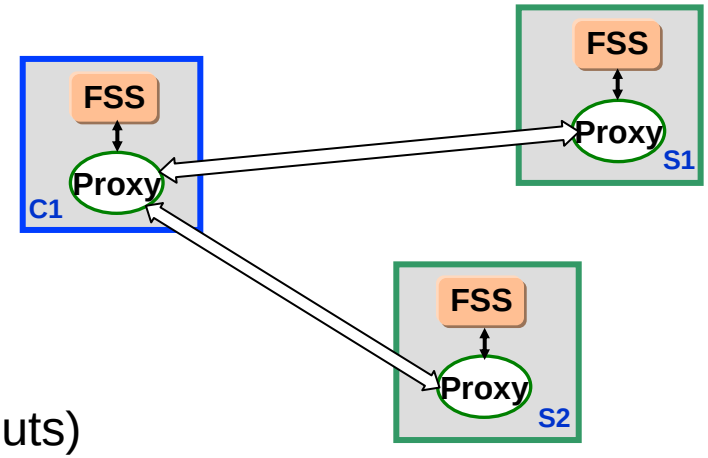


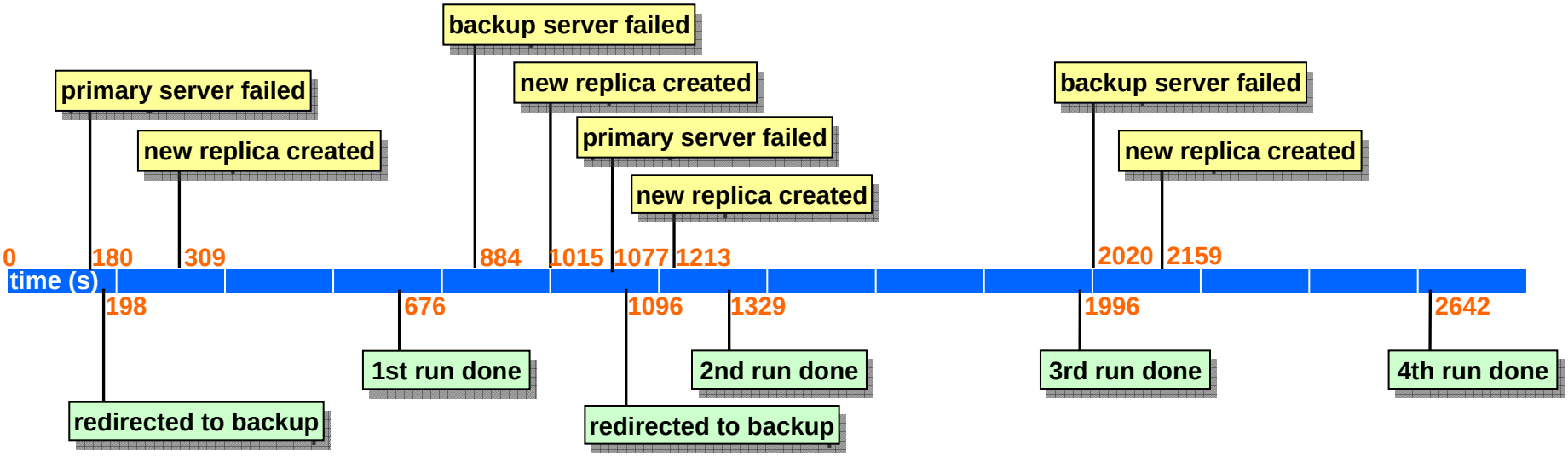


- **Autonomic session redirection achieves the best performance by adapting to the changing server load**

Autonomic Data Replication

- Scenario
  - Data transfer and replication under dynamic server failures
- Setup
  - Replica degree: 2 (1 primary + 1 backup)
  - Failures randomly injected on the servers
  - Consecutive executions of the benchmark
- Case I: ***Independent sessions*** (different inputs)

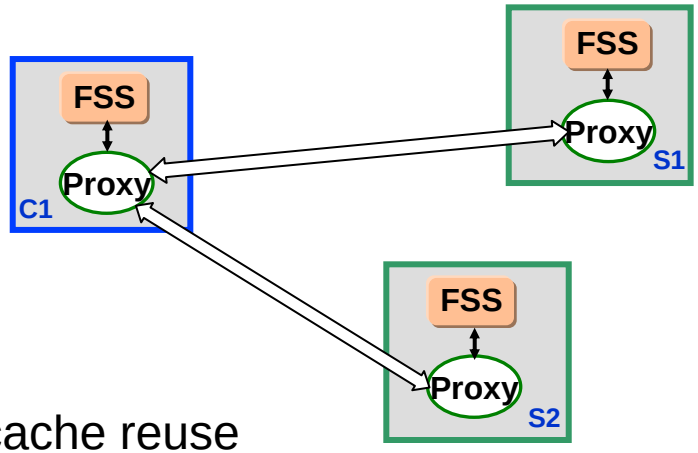




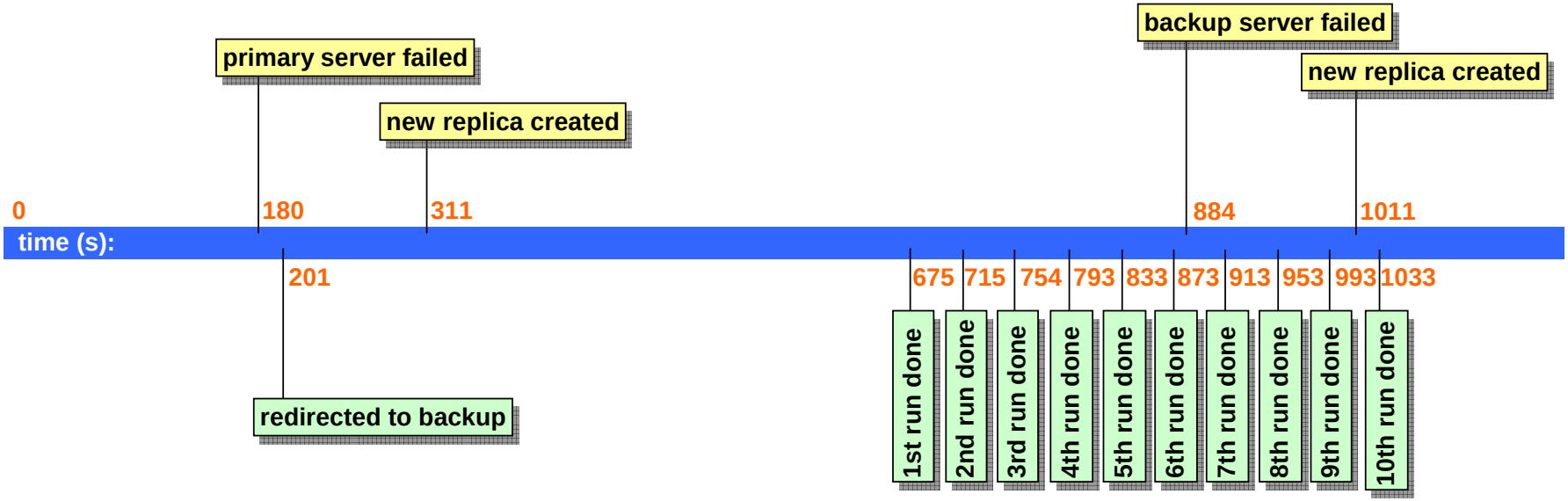
- **Autonomic data replication provides transparent error detection/recovery**

Autonomic Data Replication

- Scenario
  - Data transfer and replication under dynamic server failures
- Setup
  - Replica degree: 2 (1 primary + 1 backup)
  - Failures randomly injected on the servers
  - Consecutive executions of the benchmark



- Case II: ***Dependent sessions***: same input, cache reuse



- **Client-side disk cache further helps to minimize the impact of failures**

Related Work

- Data management in Grid environment
  - GridFTP, GASS, Condor, Legion
- Middleware control over Grid data transfers
  - Batch-Aware Distributed File System [NSDI'04]
  - Self-optimizing, fault-tolerant bulk data transfer framework based on Condor and Stork [ISPDC'03]
- Replication and storage management
  - Wide-area data replication based on Globus RFT
  - IBM autonomic storage manager

Summary

- Problem:
  - Efficient data management in heterogeneous, dynamic and large-scale Grid environments
- Solution:
  - Autonomic data management system based on GVFS and self-managing, goal-driven services
- Future work:
  - Autonomic session checkpointing and migration
  - Decentralized coordinating per-domain DSS/DRS

References

[1] M. Zhao, J. Zhang and R. Figueiredo, “Distributed File System Virtualization Techniques Supporting On-Demand Virtual Machine Environments for Grid Computing”, Cluster Computing, 2006.

[2] M. Zhao, V. Chadha and R. Figueiredo, “Supporting Application-Tailored Grid File System Sessions with WSRF-Based Services”, HPDC-2005.

[3] J. Xu, S. Adabala and J. A.B. Fortes, “Towards Autonomic Virtual Applications in the In-VIGO System”, ICAC-2005.

[4] L. W. Russell et al, “Clockwork: A New Movement in Autonomic Systems”, IBM Systems Journal, 42(1), 2003.

[5] J. Bent et al, “Explicit Control in a Batch-Aware Distributed File System”, NSDI-2004.

[6] T. Kosar et al, “A Framework for Self-optimizing, Fault-tolerant, High Performance Bulk Data Transfers in a Heterogeneous Grid Environment”, ISPDC-2003.

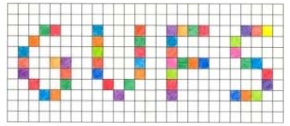
[7] A. Chervenak et al, “Wide Area Data Replication for Scientific Collaborations”, Grid-2005.

[8] Y. Zhong, S. G. Dropsho, C. Ding, “Miss Rate Prediction across All Program Inputs”, PACT-2003.

WSRF::Lite An Implementation of Web Services Resource Framework
http://www.sve.man.ac.uk/Research/AtoZ/ILCT

GVFS Virtualized distributed file system for Grid environment
http://www.acis.ufl.edu/~ming/gvfs

In-VIGO Virtualization middleware for computational Grids
http://www.acis.ufl.edu/invigo



Acknowledgments

- In-VIGO team
  - http://invigo.acis.ufl.edu

- NSF Middleware Initiative
- NSF Research Resources
- IBM Shared University Research
- Intel ISTG R&D Council

- Questions?